DOCUMENT RESUME

ED 243 894 TM 830 276

AUTHOR	Powers, Donald E.; Stansfield, Charles W.
TITLE	The Test of Spoken English as a Measure of Communicative Ability in the Health Professions: Validation and Standard Setting.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-83-1; TOEFL-RR-13
PUB DATE	Jan 83
NOTE	74p.; Some tables contain small print.
AVAILABLE FROM	TOEFL Program Office, Educational Testing Service, Princeton, NJ 08541. ($3.00; 10 or more, $2.00 each).
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Certification; Communicative Competence (Languages); Cutting Scores; *English (Second Language); *Health Personnel; *Language Proficiency; Language Tests; Oral Language; Speech Tests; Standardized Tests; *Standards; *Test Construction; *Test Validity
IDENTIFIERS	Nonnative Speakers; Test of English as a Foreign Language; *Test of Spoken English

ABSTRACT

The increasing interest in oral proficiency during the past decade prompted Educational Testing Service (ETS) to undertake the development of the Test of Spoken English (TSE), a standardized test of speaking proficiency of non-native speakers of English. The test has been validated for the selection of non-native teaching assistants applying to U.S. institutions. The objectives of the study reported here were to: (1) provide additional validation of the TSE for the selection and certification of non-native health professionals, (2) establish a set of procedures for determining standards of proficiency on the TSE, and (3) try out these procedures in selected health professions in which the test could be used. The results of the study support the validity of the TSE as a measure of oral language proficiency and provide some progress toward the setting of appropriate standards of oral language proficiency within the health-related professions. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED243894

TOEFL Research Reports

REPORT 13
JANUARY 1983

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

[X] This document has been reproduced as received from the person or organization originating it.

[ ] Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C.W. Stansfield

& H. C. Weidemiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

THE TEST OF SPOKEN ENGLISH AS A MEASURE OF COMMUNICATIVE ABILITY IN THE HEALTH PROFESSIONS: Validation and Standard Setting

Donald E. Powers
Charles W. Stansfield

TM 830 274

TOEFL
EDUCATIONAL TESTING SERVICE

The Test of English as a Foreign Language (TOEFL) was developed in 1963 by a National Council on the Testing of English as a Foreign Language, which was formed through the cooperative effort of over thirty organizations, public and private, that were concerned with testing the English proficiency of non-native speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board assumed joint responsibility for the program and in 1973 a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations (GRE) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a Policy Council that was established by, and is affiliated with, the sponsoring organizations. Members of the Policy Council represent the College Board and the GRE Board and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

A continuing program of research related to TOEFL is carried out under the direction of the TOEFL Research Committee. Its six members include representatives of the Policy Council, the TOEFL Committee of Examiners, and distinguished English-as-a-second-language specialists from the academic community. Currently the committee meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Research Committee serve three-year terms at the invitation of the Policy Council; the chair of the committee serves on the Policy Council.

Because the studies are specific to the test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. However, many projects require the cooperation of other institutions, particularly those with programs in the teaching of English as a foreign or second language. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. Local research may sometimes require access to TOEFL data. In such cases, the program may provide this data following approval by the Research Committee. All TOEFL research projects must undergo appropriate ETS review to ascertain that the confidentiality of data will be protected.

Current (1981-82) members of the TOEFL Research Committee include the following:

G. Richard Tucker (chair)	Center for Applied Linguistics
Louis A. Arena	University of Delaware
H. Douglas Brown	University of Illinois at Urbana-Champagne
Frances B. Hinofotis	University of California at Los Angeles
Diane Larsen-Freeman	The Experiment in International Living
David S. Sparks	University of Maryland

The Test of Spoken English as a Measure of Communicative Ability in the Health Professions: Validation and Standard Setting

Donald E. Powers

Charles W. Stansfield

Educational Testing Service
Princeton, New Jersey

RR-83-1

Copyright © 1983 by Educational Testing Service. All rights reserved.
Unauthorized reproduction in whole or in part is prohibited.

Table of Contents

	Page
Preface	vii
Acknowledgments	ix
Abstract	xi
Introduction	1
Relationship of Indirect Measures to Active Speaking Skills	1
The Role of Speaking Skills in Job Performance	1
Development of the Test of Spoken English	2
Description of the Test of Spoken English	3
Initial Validation of TSE	4
Methods	5
Kinds of Judgments	5
Procedures for Collecting Judgments	7
Selection of Judges	8
Data Collection	9
Data Analyses	11
Reliability of ratings	11
Regressions of ratings on TSE scores	11
Results	12
Description of Judges	12
Employment	12
Background characteristics	12
Contact with nonnative speakers	17
Additional background on consumers	17
Summary	17

Page

Reliability of the Procedures 20

Relationships of Judgments to TSE Scores 23

Summary and Discussion 45

Other Considerations 46

Allowing exceptions to the decision rule 46

Retesting speakers. 46

Using normative data to set standards on the TSE 46

Allowing changes in standards over time 47

Setting different standards for different groups 47

Next Steps 47

References 49

Appendices

A. Arrangement of TSE Standard-Setting Tapes 51

B. Percentages of Judges Rating Each Speaker as Acceptable by TSE Score Level 52

List of Tables

Table		Page
1.	Judgments Made by Each Group of Judges	6
2.	Numbers of Judges by Group and Tape	13
3.	Description of Employment Practices of Judges by Profession	14
4.	Descriptive Profile of Judges: Background Characteristics	16
5.	Descriptive Profile of Judges: Contact with Non-native Speakers	18
6.	Additional Descriptive Information for Consumer Judges	19
7.	Consistency of Judges Over Two Trials	21
8.	Intertrial Correlations for All Judges	22
9.	Logistic Regression Coefficients for Each Rating for Each Group	24
10.	Approximate Standards on the TSE Computed by Three Methods	41
11.	Cost Matrices Derived for Each Group of Judges	43

List of Figures

Figure		Page
1.	Regression of judges' ratings on TSE total comprehensibility scores. (Nurses--Hospital Nursing)	25
2.	Regression of judges' ratings on TSE total comprehensibility scores. (Nurses--Public Health)	26
3.	Regression of judges' ratings on TSE total comprehensibility scores. (Nurses--Teaching)	27
4.	Regression of judges' ratings on TSE total comprehensibility scores. (Pharmacists--Direct Contact with Public)	28
5.	Regression of judges' ratings on TSE total comprehensibility scores. (Pharmacists--Communication with Medical Colleagues)	29
6.	Regression of judges' ratings on TSE total comprehensibility scores. (Pharmacists--Teaching)	30
7.	Regression of judges' ratings on TSE total comprehensibility scores. (Physicians--Direct Care of Patients)	31
8.	Regression of judges' ratings on TSE total comprehensibility scores. (Physicians--Communication with Medical Colleagues)	32
9.	Regression of judges' ratings on TSE total comprehensibility scores. (Physicians--Communication with Public)	33
10.	Regression of judges' ratings on TSE total comprehensibility scores. (Veterinarians--Private Practice)	34
11.	Regression of judges' ratings on TSE total comprehensibility scores. (Veterinarians--Regulatory or Public Health)	35
12.	Regression of judges' ratings on TSE total comprehensibility scores. (Veterinarians--Research and Teaching)	36
13.	Regression of judges' ratings on TSE total comprehensibility scores. (Consumers--Rating of Pharmacists)	37
14.	Regression of judges' ratings on TSE total comprehensibility scores. (Consumers--Rating of Physicians)	38
15.	Regression of judges' ratings on TSE total comprehensibility scores. (Consumers--Rating of Nurses)	39
16.	Regression of judges' ratings on TSE total comprehensibility scores. (Consumers--Rating of Veterinarians)	40

Preface

Test developers have a professional responsibility not only to construct technically sound measuring instruments, but to assist test users in the proper use of these devices. The study reported here was undertaken to provide some guidance on the use of the Test of Spoken English (TSE) in several health professions in which oral communication plays a role.

One outcome of the study is a set of analyses showing the relationships between scores on the TSE and judges' assessments of nonnative speakers' facility with spoken English. Ratings were gathered from practicing health professionals in four fields for several communication situations in which medical professionals are often involved. Although these data provide potential test users with an empirical basis for determining rational standards, they do not automatically suggest definite test score cutoffs. The data are useful only when considered with other judgments, such as the importance of oral communication in various situations and the consequences of wrong decisions about both proficient and nonproficient speakers.

A second, potentially more important outcome is a set of procedures by which potential TSE users might conduct their own individually tailored standard-setting studies, employing judges whom they think qualified to make judgments and specifying communication situations that may be even more critical than those we have considered. Our hope is that certification agencies and other potential TSE users will be encouraged by the availability of these methods.

Acknowledgments

Numerous people contributed to the successful completion of this project. Among those to whom we owe thanks are:

The TOEFL Research Committee, for providing the necessary financial support, and to its individual members for useful suggestions at the inception of the project;

Ms. Cynthia Engman and Dr. Joan Large, directors of nursing education at the University of Pennsylvania and Villanova University, respectively, and Ms. Adele Herwitz, Director of the Commission on Graduates of Foreign Nursing Schools, who helped arrange data collection for nurses;

Drs. Joseph Gloyd and Leland West of the American Veterinary Medical Association, who facilitated data collection for veterinarians;

Mr. Samuel A. Kahn, Executive Director of the Greater Mercer Comprehension Planning Council, and his staff, who coordinated data collection for pharmacists, physicians, and consumers;

ETS staff members Richard Duran, Gordon Hale, Paul Angelis, Joan Borum, Samuel Livingston, Gay MacQueen, Dennis O'Toole, and Russell Webster, who contributed to various aspects of the design of the project and the review of results;

Richard Harrison, who conducted the data analyses;

Other ETS staff, including Mary Beth Brookshaw, Cathy Clauss, Frieda Hardy, Delorise Stephens, and Annette Miller, who assisted in various project activities;

Christine Sansone, Eleanor Hibbs, and Faith Thompson, who typed the final draft; and

Lorraine Simon, who coordinated numerous aspects of the project, including the preparation of the final manuscript

Finally, we thank all the anonymous nurses, pharmacists, physicians, veterinarians, and consumers who provided the ratings on which this report is based.

Abstract

The ability to communicate effectively is considered critical to successful performance in numerous occupations, including many health-related professions. Oral proficiency is necessary in some degree for performing such tasks as interacting with medical colleagues and counseling or instructing patients. The importance of communicative competence is evidenced by a growing number of states' and certification agencies' requirements for minimum oral proficiency in a variety of occupations.

The increasing interest in oral proficiency during the past decade prompted Educational Testing Service (ETS) to undertake the development of the Test of Spoken English, a standardized test of speaking proficiency of nonnative speakers of English. The test has been validated for the selection of nonnative teaching assistants applying to U.S. institutions.

The objectives of the study reported here were to: (a) provide additional validation of the TSE for the selection and certification of nonnative health professionals, (b) establish a set of procedures for determining standards of proficiency on the TSE, and (c) try out these procedures in selected health professions in which the test could be used.

Groups of judges who were practitioners in four health professions--nursing, pharmacy, medicine, and veterinary medicine--judged the acceptability of the speaking performance of examinees who had taken the TSE. Each group was asked to provide, for its profession, ratings for three distinct situations, such as communicating with professional colleagues, teaching in the particular field, or working in a hospital. Judges were asked to indicate whether each examinee was at least minimally acceptable to function in each situation. Judges' ratings were then related to the examinees TSE scores. A larger group of consumers of medical care gave more global ratings to the same examinees for each of the four professions.

Background data suggested that the groups of judges were heterogeneous with regard to a number of relevant personal characteristics and experiences, such as degree of exposure to nonnative speakers. This diversity suggests that judges represented a variety of points of view about necessary language proficiency.

Analyses of judges' ratings revealed a substantial degree of consistency. Even though only relatively small segments of test performance were rated, judges tended to assign the same ratings when asked to rate a second sample of test performance for the same examinees. Although the data showed that the procedures applied can sometimes produce different results from one occasion to the next, the relationships of judges' ratings to TSE scores generally exhibited considerable stability.

Alternative methods were applied to determine possible ranges of cutoff scores on the TSE. One of these methods involved a consideration of the consequences of different decision outcomes, e.g., whether the certification or licensing of an unacceptable speaker is more or less serious than failing to certify an acceptable speaker. A range of possible cutoff scores was computed for each situation within each profession.

The results of the study support the validity of the TSE as a measure of oral language proficiency and provide some progress toward the setting of appropriate standards of oral language proficiency within the health-related professions.

Introduction

Relationship of Indirect Measures to Active Speaking Skills

Through most of this century, the multiple-choice test has been the primary mode of assessing proficiency in English as a second language. Most tests currently in use--such as the Test of English as a Foreign Language (TOEFL), the Michigan Test of English Language Proficiency (MTELP), the Comprehensive English Language Test (CELT) and the American Language Institute of Georgetown University (ALIGU) tests--usually employ a multiple-choice format. Such tests typically measure listening comprehension, reading comprehension, vocabulary, and English structure. The more active skills, such as writing, generally have been measured indirectly through such test item types as error recognition.

None of these tests actually measures oral proficiency, though some research suggests that the relationship between multiple-choice indirect measures and direct measures of active speaking skills is moderately high. Studies by Carroll (1967), Clark and Swinton (1980), Osanyinbi (1974), and Pike (1979) show moderate correlations between scores on indirect, paper-and-pencil measures of language proficiency and performance on tests of active speaking skills. Not surprisingly, listening comprehension shows a stronger relationship with speaking skills than does reading comprehension or writing ability. Although the correlations are significant, an individual's score on one measure will not necessarily be a very good predictor of his or her score on another measure, because of individual differences in such factors as language learning history and aptitude.

For instance, individuals receiving nearly all their exposure to English in a formal classroom setting might exhibit considerably greater facility in reading and writing than in listening and speaking. In contrast, persons whose principal exposure to English has been more informal might exhibit an entirely different profile of acquired skills. Similarly, individual variation in aptitude for learning a second language, in cognitive style, in age, in degree of ethnocentricity or acculturation, and in purpose for learning the language (Schumann, 1976) will also affect the learner's profile of language skills. Therefore, indirect, paper-and-pencil measures are of somewhat limited value in assessing the oral proficiency of individual examinees. When drawing inferences about active speaking skills, then, there would appear to be no acceptable substitute for a valid direct measure of such skills.

The Role of Speaking Skills in Job Performance

In many occupations and professions, the ability to communicate effectively is considered critical to successful performance on the job. Recognizing this, the Foreign Service Institute (FSI) developed the FSI absolute proficiency rating scale during the 1950s (Sollenberger, 1978). The scale, which provides a rating based on a face-to-face interview, is

used to determine individuals' suitability for particular jobs requiring proficiency in the spoken language, e.g., those positions held by US Foreign Service officers at overseas embassies and consular posts.* This scale is also used by the Defense Language Institute, the Central Intelligence Agency, and the National Security Agency to certify the English language proficiency of personnel in various job assignments. The Peace Corps considers scores on an FSI-type oral interview in deciding promotions of trainees to volunteer status.

Within the context of bilingual education, and the teaching of English as a second language, measures of active speaking have been developed by individual school districts and state education agencies. Some states, such as Texas and New Jersey, use a modified FSI-type interview to grant teaching credentials in these fields.

Within the broader context of teaching in general, language proficiency is also held to be essential to job performance. At a growing number of universities, proficiency in spoken English is required of nonnative instructional staff; in Oklahoma, all alien instructors within the state system of higher education are required to demonstrate proficiency in spoken English.

At the Department of Regulation and Licensing of the state of Wisconsin, Showers and Nyenhuis (1982) have called attention to the importance of oral proficiency in nursing. They found a correlation of .63 between the oral English proficiency of foreign-born nurses and supervisors' ratings of job performance. Successful performance of such tasks as gathering information about patients, discussing patient needs, counseling, motivating and teaching patients, and recording patients' responses to treatment depends to a considerable degree on oral proficiency. In recognition of this dependence, Wisconsin requires that all foreign-trained nurses pass an oral English proficiency test.

Development of the Test of Spoken English

Responding to growing interest in oral English proficiency testing during the past decade, in 1978 Educational Testing Service undertook the development of the Test of Spoken English. Anticipated applications were in employment situations where oral skills are considered critical (Clark & Swinton, 1979). A factor in test design considerations was the need to elicit evidence of general speaking proficiency rather than proficiency in a particular language-use situation. Because the instrument might be used to determine examinee suitability in several contexts, the test could not contain item formats or individual questions that would require extensive familiarity with a particular subject matter area or employment context.

---

*The FSI interview is now also known as the Interagency Language Round Table (ILR) Oral Proficiency Interview.

Following an examination of a variety of question types used in previously developed "semidirect" speaking tests, 11 separate formats and question types were selected for inclusion in an experimental test. These included item types eliciting discrete language elements from the examinee (for example, particular vocabulary items and morphological or syntactic patterns), as well as general formats that allowed more freedom in responding.

The form was administered experimentally and examinee performance on each item type was related to performance on both TOEFL and the FSI oral proficiency interview. Item types and specific items in the operational test form were selected with the goal of maximizing the correlation with the FSI and minimizing the correlation with TOEFL score. Correlations between the TSE prototype items and the separate FSI ratings for fluency, pronunciation, comprehension, grammar, and vocabulary were also examined with a view toward deleting items that correlated more strongly with the grammar and vocabulary ratings, or with scores on Sections II and III of TOEFL, which involve written language only.

Description of the Test of Spoken English

The TSE consists of seven sections, each involving a particular speech activity. The first section is an unscored "warm-up" in which examinees respond orally to a short series of biographical questions spoken on the test tape (name, reasons for studying English, future plans, etc.). In the second section, examinees read aloud a short printed passage (approximately 125 words) and are told they will be scored for proper pronunciation and overall clarity of speech. Time is allowed for preliminary silent reading of the passage.

In the third section, examinees see a series of ten partial sentences and are asked to complete each sentence orally in a way that conveys meaning and is grammatically correct. The fourth section consists of six line drawings that "tell a continuous story." After studying the drawings briefly, examinees are asked to tell the story that the pictures show, narrating in the past tense. In the fifth section, examinees looks at a single line drawing and answer a series of spoken questions about the picture's content.

The sixth section consists of a series of spoken questions intended to elicit free and somewhat lengthy responses from the examinees. The questions require both descriptions of common objects and open-ended expressions of opinion on familiar issues. The linguistic quality and adequacy of communication rather than the specific knowledge revealed in an examinee's response are considered in scoring this section. In the seventh and final section, examinees see a printed schedule, such as a schedule for an imaginary college course, and are asked to describe the schedule aloud, as though informing a group of students on the first day of class.

Tape recordings of TSE examinees' answers are rated by experienced teachers who are trained specialists in the field of English as a second language. Raters are trained at one-day workshops conducted by ETS staff and are retrained annually (or more frequently, if score discrepancies indicate the need for such retraining).

Currently, procedures require that examinee response tapes be rated independently by two TSE raters. If disagreements occur, a third rater is engaged to resolve the discrepancies.

Scores in each TSE diagnostic area (pronunciation, fluency, and grammar) are reported on a scale ranging from 0 to 3. The overall comprehensibility score is reported on a scale ranging from 0 to 300.

Initial Validation of TSE

To provide initial validation of the TSE, Clark and Swinton (1980) administered both the TSE and the FSI interview to 60 instructors (foreign teaching assistants) at nine large state universities. Scores were correlated with student ratings of the instructors' use of language in both the classroom and other instructional contexts, and with a standardized instructor/course evaluation instrument. The observed correlation between TSE and FSI interview scores was .79. Multiple regression analyses showed that both TSE and FSI served as good predictors of student ratings of instructors' communicative proficiency in specific instructional tasks as well as in more general aspects of teaching performance. TSE scores correlated with rated proficiency in delivering classroom lectures (.60), understanding student questions (.52), answering questions (.53), and communication during office appointments (.54). They also correlated (.68) with the degree to which the instructor's pronunciation interfered with understanding. To a lesser extent, TSE scores correlated (.29) with the instructor's "overall teaching effectiveness" and with other nonlinguistic variables, such as interpersonal relations.

The high correlation between TSE and FSI scores would seem to justify use of the TSE in situations where it is impractical to conduct face-to-face interviews. The strong correlations of TSE with student perceptions of instructors' proficiency indicate that the test can predict such proficiency to an acceptable degree.

The objectives of the study described here were to (a) provide additional validation of the TSE for the selection or certification of nonnative health professionals, (b) establish a set of procedures for determining standards of proficiency on the TSE, and (c) try out these procedures in selected health professions in which the test could be used. Briefly, objective (a) was to be accomplished by enlisting various groups of health professionals and consumers of medical services to rate the oral proficiency of a sample of TSE test takers. These ratings would then be related to actual test scores. The establishment of procedures under objective (b) involved sampling and arranging the test performances on

which judgments were to be made, developing the instructions to raters, and specifying the methods by which judgments would be related to test scores. Objective (c) would entail the actual implementation and evaluation of these procedures.

Methods

The procedures used were modeled generally after those employed in a study by Livingston (1978), which was designed to set standards of speaking proficiency for teachers of English as a second language and for teachers of Spanish-English bilingual classes. For the study described here, the following specific procedures were followed.

Kinds of Judgments

The first task was to decide on the nature of the judgments to be made, i.e., the situations to be considered by judges. We decided that asking for ratings of more than three situations would not be feasible. Hence, each group of health professionals--nurses, pharmacists, physicians, and veterinarians--was asked to rate three distinct situations thought to represent the kinds of activities in which these professionals might be engaged. These situations are listed in Table 1, along with examples of the possible characteristics of each situation. Consumers were asked to rate only a particular pair of the six possible pairs of the four professions.

Of necessity, we asked for relatively global ratings of situations that would represent the various kinds of situations in which each of the health professionals might be involved. No attempt was made to be comprehensive, nor was effort directed toward very specific situations. For example, physicians were asked to judge oral proficiency needed in the "direct care of patients." We might have asked about the degree of speaking proficiency needed in the performance of surgical procedures, for example (in which oral proficiency might be even more critical), but time limitations precluded such detail. Likewise, although some clinical specialties (e.g., surgery, pediatrics, psychiatry) may require different degrees of oral proficiency than others, no attempt was made to achieve this level of detail. In addition, it was not thought practical within the limits of this study to consider other important dimensions of communicative competence (e.g., interpersonal skills and various affective components) or functions of language (e.g., persuasion or developing rapport) that might be highly desirable in the professions under consideration.

The situations that were finally included were selected for a variety of reasons. Some (e.g., those used for physicians) had been mentioned as important in the publications of such agencies as the Educational Commission on Foreign Medical Graduates (ECFMG). Others were suggested by

Table 1

Judgments Made by Each Group of Judges

Judges	Situation	Examples of Characteristics of Each Situation
Nurses	(1) Hospital nursing	Providing bedside nursing care and carrying out prescribed medical treatment plans.
	(2) Public health nursing	Instructing patients and families in proper health care in clinics, homes, schools, and other community settings. Working with physicians and other community leaders in health education.
	(3) Teaching	Teaching the principles of nursing in a classroom setting or in the direct care of patients. Conducting continuing education classes.
Pharmacists	(1) Direct contact with the general public	Dispensing prescriptions and offering advice on their use, usually in community pharmacies.
	(2) Communication with medical colleagues	Advising hospital staff on the selection and effects of drugs. Explaining the technical aspects of drugs to physicians and pharmacists.
	(3) Teaching	Teaching at colleges of pharmacy or other educational institutions.
Physicians	(1) Direct care of patients	Communicating with patients and with their families.
	(2) Communication with medical colleagues	Interacting with colleagues in institutional settings or in professional organizations.
	(3) Communication with the general public	Interacting with governmental and non-governmental agencies and with the public media.
Veterinarians	(1) Private practice	Conducting a general practice such as those dealing with farm animals or companion animals.
	(2) Regulatory medicine or public health	Working in governmental and non-governmental agencies to inspect meat, poultry, and livestock and to control animal disease.
	(3) Research and teaching	Holding academic positions in agricultural schools or in veterinary colleges.
Consumers	(1) Nurse	In hospitals, clinics, schools, or other community settings.
	(2) Pharmacist	Dispensing medications in community pharmacies.
	(3) Physician	In hospitals, clinics, schools, or other community settings.
	(4) Veterinarian	In clinics, animal hospitals, or other settings.

descriptions of job requirements and qualifications, working conditions, nature of work, and places of employment outlined in such publications as the Encyclopedia of Careers and Vocational Guidance and the Occupational Outlook Handbook. A formal task analysis of each situation, though desirable, was well beyond the scope of the study, however.

Procedures for Collecting Judgments

The stimuli on which judgments were made were actual samples of TSE test performances. All persons (N = 486) who had taken the test in either of the first two TSE administrations (October 1979 or March 1980) were stratified by TSE score level (25-point intervals) and native language. Although the entire range (0-300) of TSE scores was represented, performances were sampled in greater numbers in the middle of the score range to maximize measurement precision, since minimum proficiency levels were expected to lie in this intermediate range. Only the seven language groups (African, Arabic, Chinese, Farsi, Greek, Japanese, and Spanish) that were most heavily represented in these administrations were used. A total of 36 protocols (approximately equal numbers from each language group) was eventually selected on these bases. TOEFL program staff then judged these tapes for clarity, i.e., the absence of any background noise and other distortion that might distract judges unnecessarily. Although audible, three tapes were judged to be somewhat lower in quality than the others and were subsequently replaced by other tapes randomly selected from the same score levels and language groups. The distribution of scores on each tape is shown in Appendix A.

Because each TSE test protocol is about 20 minutes long, test performances were sampled for each of the 36 selected protocols so judges could listen to a greater number of examinees in the limited time available. Accordingly, only three of the six operational test sections (II, VI, and VII) were used. These sections differ in the kinds of responses required and thus were thought to provide a good representation of the total test. For Section II the examinee is asked to read a printed paragraph silently and then to read it aloud. In Section VI the examinee is required to give opinions on topics of international interest and to describe certain objects. Section VII entails explaining a class schedule upon meeting a class of students for the first time. For Sections II and VII, examinees' responses to all questions were included on the tapes. For Section VI, the responses to two of the three questions were used, thus resulting in one to one and a half minutes of speech from each of the three sections. A comparison of subscores for each section did not reveal any systematic differences between these three sections and the three that were not considered in this study. Thus, no unintentional bias appears to have been introduced by the sampling of test sections.

The 36 test protocols, stratified on score level and native language, were assigned to three equal groups, each of which was placed on a separate tape. Two test sections were chosen to represent each protocol--either II and VI, II and VII, or VI and VII--thus providing a total of two to three minutes of speech for each examinee protocol.

As shown in Appendix A, protocols were positioned in a random order in the first half (12 positions) of each tape. Equal numbers of examinees represented each of the sections used (II, VI, and VII). In the second half of the tape the same 12 examinees were included, again in a random order, but with a different section of the test. Again, each section was represented equally in this half of the tape. The order of appearance of the sections was completely counterbalanced, with each section appearing equally often as first or second. To allow the judges to "warm up," three unanalyzed practice protocols, representing low, medium, and high TSE scores, were placed at the beginning of each tape.

Each test performance was preceded by the directions given to test takers. The complete, unabridged instructions preceded the practice protocols, but in the interest of efficiency only much briefer, edited directions were included with all the protocols that followed.

Each group of judges was asked to provide ratings of the acceptability of each of 27 protocols. (Judges were not informed that the first three protocols were for practice only, nor were they told that the remaining 24 protocols were actually only 12 examinees appearing in each half of the tape.)

Selection of Judges

Each group of judges was selected in a slightly different manner. For veterinarians, we worked through the Educational Commission for Foreign Veterinary Graduates (ECFVG) of the American Veterinary Medical Association (AVMA). A letter was sent by the AVMA's Director of Scientific Affairs to each of 50 alternate state delegates to the AVMA 1981 annual convention. These delegates were reminded that most states require graduates of veterinary colleges located outside the United States and Canada to obtain a certificate from the ECFVG before taking state licensure examinations and that, in order to obtain an ECFVG certificate, candidates must demonstrate their ability to communicate in spoken English by taking the TSE. The delegates were also informed of the purpose of the study (to gather information to help determine appropriate standards of proficiency on the TSE).

For nurses, we approached the director of the Commission on Graduates of Foreign Nursing Schools (CGFNS), who in turn contacted directors of nursing at two graduate nursing schools in the Philadelphia area--one, Villanova University, located in the Philadelphia suburbs and the other, the University of Pennsylvania, located in central Philadelphia. Again, the objectives and study procedures were explained, and the directors were asked to assist in recruiting participants. We left the specific procedures of selecting judges to the directors of nursing, who were asked only to provide diverse samples of nurses who might be involved in different activities or clinical specialties and who may have had varying degrees of exposure to nonnative speakers of English.

The participation of pharmacists, physicians, and consumers was arranged by the Executive Director of the Greater Mercer Comprehensive Planning Council (GMCPC), which coordinates health care planning in Mercer County, N.J. (Mercer County is located in central New Jersey and includes many kinds of rural, suburban, and urban communities, including the town of Princeton and the city of Trenton.) The GMCPC is comprised of both providers and consumers of health care services. Its individual members represent a cross-section of Mercer County residents with respect to socioeconomic status, educational level, sex, ethnicity, and geographic location within the county.

The Executive Director was asked to recruit pharmacist-, physician-, and consumer-volunteers from the GMCPC membership so participants would reflect the diversity within the Council. In addition, physicians were selected so as to represent a variety of clinical specialties.

Data Collection

The standard-setting sessions were group administered. At the start of each session judges were given the following:

(a) an explanation of the objectives of the study (i.e., to help determine appropriate standards of proficiency on the TSE)

(b) background about the development and purpose of the TSE

(c) information about the potential application of the TSE in the certification of foreign-trained health professionals

The various situations that participants were to rate were also discussed. Judges were then asked to listen to each of the protocols and to record whether the proficiency of the speaker was at least minimally sufficient for the individual to be able to function adequately in each of the situations. Judges were also given the following instructions.

1. You are being asked to judge speaking proficiency only, not academic background, experience, or any other factors. When listening to the interviews, you should disregard the particular topic being discussed and listen only for the degree of speaking proficiency shown by the speaker.

2. Your judgment should be based upon your acquaintance with the kind of speaking ability and the amount of speaking that is called for in each of the situations listed. Please try to think of these situations in terms of your own personal experiences.

3. Please keep constantly in mind that the judgment being asked for is that of minimally acceptable speaking proficiency, not ideal or optimum proficiency. In listening to each

speaker, ask yourself if the person could function in a just-adequate way with respect to speaking proficiency. If the person's speech is at least minimally acceptable, please make a check under "acceptable"; if not, make your check under "not acceptable." The placement of your checkmark should also indicate your degree of certainty in your judgment. For example, speakers that you judge to be just barely acceptable (or not quite acceptable) would receive checkmarks near the line that separates the "not acceptable" and "acceptable" categories. Ratings indicating completely acceptable would be made at the far right of the scale, and those for not at all acceptable at the far left. (A "barely acceptable" rating would be indicated as follows: |_____|✓_____|).
Not Acceptable Acceptable

4. Please try to make independent judgments of each speaker's acceptability for each of the situations. The standards that you apply to each situation may differ or they may be the same, depending on how you perceive the speaking requirements of each situation.

5. The samples of speech will be played in a random order. Because the decision is up to you, there is no predetermined number of acceptable and not acceptable speakers. Do not be concerned if you find that you are marking a large number, or even all, of the speakers as acceptable or not acceptable.

6. You may find that some speakers are more difficult to judge than others. However, please make a judgment for each speaker, even if the choice is difficult. For speakers that may be quite easy to judge, we recommend that you listen to the speaker's entire speech before making your response.

7. For each speaker an announcer will indicate the question to which the speaker was asked to respond and the number of the speaker. At the end of each speaker's response you will be asked to record your judgments for the speaker. Please do not discuss your judgments with your colleagues until all of the judging is over.

Judges were divided into subgroups, with each subgroup judging a different sample of examinees. Some groups provided enough representatives to require three or more sessions--at least one for each sample of examinees (i.e., each tape). Other groups had large enough numbers for only two sessions, hence only two tapes.

In order to describe the judges, separate questionnaires were developed and administered to each group of judges to ascertain:

(a) the amount of contact with nonnative speakers

(b) type of occupation (for consumers), practice, or medical specialty

(c) type and place of employment

(d) age and sex

(e) geographic location

(f) language background

Judges were also asked to rate the seriousness of making each of two types of errors in decisions about granting licenses in their professions: (a) licensing a person who is not a minimally acceptable speaker and (b) failing to license a speaker who is minimally acceptable.

Data Analyses

Although the forms on which judges were asked to record their ratings provided only for a distinction between acceptable and not acceptable, judges were asked to indicate the degree of certainty in their ratings by the placement of their checkmarks (see instruction #3 above). These ratings were subsequently coded on a four-point scale, with 1 and 2 used for degrees of unacceptability and 3 and 4 for degrees of acceptability, so that additional, finer-grained analyses might be undertaken.

Reliability of ratings. Because judges provided two ratings for each of 12 protocols, it was possible to estimate the consistency with which each judge assigned ratings. Product-moment correlation coefficients were computed between the two sets of 12 ratings (trial 1 vs. trial 2). Cross-tabulations between the ratings on trial 1 and those on trial 2 were also computed for each judge to ascertain the consistency of the particular decisions the judges were asked to make, i.e., to classify each of the 12 protocols as either acceptable or unacceptable.

Regressions of ratings on TSE scores. For each group of judges, ratings were regressed on TSE overall comprehensibility scores for each trial. Logistic regression functions were fitted to predict the dependent variable--the proportion of judges that considered a protocol to be acceptable. The logistic function is generally considered to provide a better fit than linear functions to data that involve proportions.

Results

Description of Judges

Table 2 shows by group the total numbers of judges upon which the results are based and the numbers who listened to each of the three tapes. The smallest groups of judges (pharmacists and physicians, N = 12 and N = 13) listened to only two of the tapes, while each of the other larger groups listened to all three. No individual judge listened to more than one tape.

Employment. Table 3 gives the types of employment in which judges (except consumers) were involved. Nurses were about equally divided among hospital nurses, teachers, and others (including nurse midwife, education coordinator, and student nurse). None of the 27 nurse judges was engaged in public health. Slightly over half (53.8%) were affiliated with a university or college, and about a third (34.6%) were employed by a hospital.

Nine of the 12 pharmacist judges listed practice in a community pharmacy as a primary responsibility. A few also had either primary or secondary duties in hospitals or clinics or as teachers or researchers. A majority (58.3%) worked in community pharmacies.

All but one of the physicians were engaged in private practice, with a wide range of specialties that included hematology, immunology, internal medicine, opthamology, obstetrics, pediatrics, and surgery. Nine of the 13 were self-employed, others were employed by health maintenance organizations, by a university hospital, or by some other organization.

The largest number of veterinarians (42.9%) maintained their own private practices. Eight others (28.6%) were involved in university-based research or teaching. Five listed involvement in other activities, which were usually administrative positions with the AVMA, and three held positions in public health or regulatory medicine.

Consumers listed a wide range of employment, including such occupations as college administrator, railroad worker, journalist, homemaker, attorney, security guard, and student.

Background characteristics. Table 4 contains information on selected background characteristics of judges. Information on these variables was thought to be potentially relevant to judges' attitudes toward nonnative speakers of English.

Although nearly all the judges were native born, a majority of each group (ranging from 58.3% of pharmacists to 84.6% of physicians) had traveled abroad at some time. Most had studied a foreign language--either in high school (ranging from half of the veterinarians to all of the pharmacists), in college (ranging from a third of the pharmacists and veterinarians to 84.6% of the physicians), or in some other setting

Table 2

Numbers of Judges by Group and Tape

Group	Tape Number 1	2	3	Total
Nurses	14	8	5	27
Pharmacists	--	6	6	12
Physicians	--	6	7	13
Veterinarians	8	11	11	30
Consumers	13	26	12	51

Note. All groups except consumers provided ratings of three situations. Consumers were divided into six groups each of which rated a pair of professions.

Table 3

Description of Employment Practices of Judges by Profession

Type of Practice or Activities	N	%
Nurses (N = 27)		
Hospital nursing	9	34.6
Public health nursing	0	0.0
Teaching	10	38.5
Other	7	26.9

	Primary Duty		Secondary Duty	
	N	%	N	%
Pharmacists (N = 12)				
Practice in a community pharmacy	9	75.0	1	8.3
Practice in a hospital or clinic	1	9.1	1	9.1
Teaching or research	1	9.1	2	18.2
Representing a pharmaceutical firm	0	0.0	0	0.0
Other	1	9.1	0	0.0
Physicians (N = 13)				
Private practice	12	92.3	0	0.0
Public health	3	23.1	1	7.7
Teaching or research	1	7.7	7	53.8
Other	2	15.4	3	23.1

	N	%
Veterinarians (N = 30)		
Private practice	12	42.9
Regulatory medicine or public health	3	10.7
Research or teaching	8	28.6
Other	5	17.9

Table 3 (Cont'd.)

Category of Employment	N	%
Nurses		
Employed by a general or a specialized hospital	9	34.6
Employed by a governmental or other public agency	1	3.8
Employed by a commercial firm or in private industry	0	0.0
Employed by a university or college	14	53.8
Other	2	7.7
Pharmacists		
Employed by (or own) a community pharmacy	7	58.3
Employed by a governmental or public agency	1	8.3
Employed by a commercial firm or in private industry	1	8.3
Employed by a university or college	0	0.0
Other	3	25.0
Physicians		
Self-employed	9	69.2
Employed by a governmental or public agency	1	7.7
Employed by a commercial firm or in private industry	2	15.4
Employed by a university or college	1	7.7
Other	4	30.8
Veterinarians		
Self-employed	12	41.4
Employed by a governmental or public agency	2	6.9
Employed by a commercial firm or in private industry	2	6.9
Employed by a university or college	8	27.6
Other	5	17.2

Note. Some responses do not total 100% because of multiple responses, and for some questions the number of responses is less than the total number of judges because of nonresponse.

Table 4

Descriptive Profile of Judges: Background Characteristics

Background Characteristic		Group				
		Nurses (N = 27)	Pharmacists (N = 12)	Physicians (N = 13)	Veterinarians (N = 30)	Consumers (N = 51)
Country of birth (% U.S.A)		96.2	91.7	92.3	96.7	100.0
Traveled abroad (%)		74.1	58.3	84.6	80.0	70.6
Studied a foreign language (%)						
In high school		85.2	100.0	92.3	50.0	80.4
In college		55.6	33.3	84.6	33.3	41.2
Other		18.5	8.3	92.3	16.7	7.8
Proficient in any foreign language (%)						
Reading		33.3	33.3	69.2	23.3	25.5
Speaking		25.9	25.0	76.9	10.0	21.6
Writing		11.1	8.3	76.9	6.7	11.8
Geographic location (%)						
Large city (over 250 thousand)		37.0	0.0	30.8	13.8	8.0
Medium-sized city (50 - 250 thousand)		48.1	66.7	61.5	44.8	52.0
Small city (10 - 50 thousand)		7.4	33.3	7.7	31.0	36.0
Rural area or town (less than 10 thousand)		7.4	0.0	0.0	10.3	4.0
Age	M	34.0	45.3	43.8	54.1	41.4
	SD	10.3	12.6	15.1	8.9	14.2
	Range	20 - 54	25 - 73	31 - [illegible]	40 - 74	18 - 76
Sex (% male)		0.0	66.7	84.6	100.0	47.1

(ranging from 7.8% of the consumers to 92.3% of the physicians). Physicians were more likely than other judges to rate themselves as proficient in another language; over two-thirds indicated proficiency in reading, speaking, and writing some foreign language. In contrast, a third or fewer judges in each of the other groups said they were proficient in any of these aspects.

The groups of judges also differed with respect to geographic location. The largest proportion of each group, however, indicated a moderate-sized city or suburb as a place of residence or employment. Because of the way data were collected, veterinarians were the most geographically diverse group, representing 20 states.

Within each group a range of ages was noted. On the average, nurses were the youngest group (M = 34.0 years, SD = 10.3) and veterinarians the oldest (M = 54.1 years, SD = 8.9). All of the veterinarians and most of the pharmacists (66.7%) and physicians (84.6%) were men; all of the nurses were women. Consumers were nearly equally divided between men and women.

Contact with nonnative speakers. One relevant characteristic of judges was thought to be the degree of contact with nonnative speakers. Table 5 suggests that judges in each group had moderately frequent contact with nonnative speakers, as friends, colleagues, clients, or otherwise. Physicians reported more frequent contact, especially with colleagues and clients, than did other groups. Within each group, however, there was considerable variation.

Physicians also reported a greater number of weekly contacts with nonnative speakers when compared with other judges. A sizable majority (84.6%) reported more than five such contacts per week, while a third or fewer of each of the other groups reported this many contacts. A minority of judges in each group said they usually had no contact with nonnative speakers in a typical week (none of the physicians to 30.0% of the veterinarians). The modal contact of judges tended to be one to ten minutes in duration. The frequency of such contact ranged from 48.1% for nurses to 70.0% for pharmacists.

Additional background on consumers. Table 6 gives some additional background on consumer judges. These data show that this group was relatively well educated on average (more than 60% had completed college or gone beyond this level), predominantly White, and fluent in English. Most of these consumers reported having professional contact with a physician (88.2%), a nurse (74.5%), or a pharmacist (68.6%) during the past year, but relatively few (33.3%) had consulted a veterinarian.

Summary. Judges' descriptive profiles suggest that each group of judges was relatively heterogeneous with respect to a number of general personal characteristics and also in terms of specific characteristics that might moderate attitudes about, or ability to comprehend, nonnative speakers, and hence make valid judgments about the oral proficiency of such speakers. Among the characteristics on which judges varied were

Table 5

Descriptive Profile of Judges: Contact with Non-native Speakers

Characteristic		Group				
		Nurses (N = 27)	Pharmacists (N = 12)	Physicians (N = 13)	Veterinarians (N = 30)	Consumers (N = 51)
Frequency of contact with non-native speakers[1]						
Friends or social acquaintances	M	2.62	1.33	2.38	2.33	2.76
	SD	1.33	.62	1.21	1.22	1.42
Colleagues or business acquaintances	M	2.42	2.58	4.54	2.47	3.04
	SD	1.25	1.55	0.84	1.20	1.63
Clients	M	2.67	2.67	3.92	1.76	2.52
	SD	1.22	1.49	1.21	1.11	1.56
Others	M	3.11	1.60	3.50	2.38	2.38
	SD	1.45	.80	0.50	1.58	1.60
Number of weekly contacts with non-native speakers						
None		11.1	16.7	0.0	30.0	11.8
1 - 2		51.9	33.3	7.7	30.0	39.2
3 - 5		25.9	16.7	7.7	26.7	21.6
More than 5		11.1	33.3	84.6	13.3	27.5
Length of typical encounter with non-native speakers						
Less than 1 minute		7.4	20.0	0.0	13.3	10.2
About 1 - 10 minutes		48.1	70.0	61.5	66.7	51.0
About 10 - 60 minutes		25.9	0.0	30.8	16.7	20.4
One hour or more		18.5	10.0	7.7	3.3	18.4

[1]Responses are on a scale of 1 = very infrequent (several times a year or less) to 5 = very frequent (daily or almost daily).

Table 6

Additional Descriptive Information for Consumer Judges

Background Characteristic	N	%
Level of education		
Less than high school	0	0.0
High school graduate	5	9.8
Some college	15	29.4
Graduate of 4-year college	3	5.9
Some work beyond college	14	27.5
Graduate or professional degree	14	27.5
Self-description		
Black, Afro-American, or Negro	9	17.6
Mexican-American or Chicano	0	0.0
Oriental or Asian-American	0	0.0
Puerto Rican	1	2.0
Other Hispanic or Latin-American	0	0.0
White	41	80.4
Other	0	0.0
Best language		
English	50	98.0
Other	1	2.0
Contact with medical personnel in past year		
Physician	45	88.2
Nurse	38	74.5
Pharmacist	35	68.6
Veterinarian	17	33.3

their involvement in the particular kinds of situations (e.g., public health nursing) they were asked to consider. Judges also differed, both within and among groups, with regard to such potentially important characteristics as their exposure to nonnative speakers of English. In addition, consumers also varied with respect to recent contact with the medical personnel they were asked to consider when making ratings.

Reliability of the Procedures

Because each judge rated each of the 12 test protocols twice, it was possible to assess both the consistency of judges' responses and the reliability of the standard-setting process. Several factors may have affected consistency, including those that contribute to both systematic errors (e.g., practice effects and fatigue) and random errors associated with human judgments. Several kinds of sampling error were involved in the standard-setting process used here. Among the sources of error were:

(1) the sampling of test sections upon which judgments were made

(2) the sampling of examinees

(3) the sampling of judges

(4) the times of judging (i.e., trials)

Any or all of these factors, either separately or in combination, may have contributed to reducing the generalizability of the study findings.

Table 7, which depicts the consistency of ratings for each group of judges over two trials, shows that judges tended to rate some protocols as acceptable on one trial but not on another. As mentioned above, some of these apparent inconsistencies may have resulted from using different sections of the test on each trial. Judges may have been correct in assigning an acceptable rating to an examinee's "reading aloud" performance from Section II of the test, while giving an unacceptable rating to the same examinee's "detailed descriptions" from Section VI. Therefore, the percentages given in Table 7 may be viewed more appropriately as lower-bound estimates of judges' consistency. Only for veterinarians did the total percent of consistent judgments fall below 70% (69.7% for rating 1 and 67.3% for rating 3). Thus, overall, judges exhibited a substantial degree of consistency.

Table 8 gives the mean ratings (expressed on a 1-4 scale with 1 and 2 as unacceptable and 3 and 4 as acceptable) and the correlations between trials for all judges. Mean ratings appear comparable over trials, and the correlations are moderately high (.40-.45 for veterinarians to .56-.65 for pharmacists). While these correlations are not remarkably high, they are impressive in light of the task involved. It should be recalled that test protocols were sampled most heavily in the score range where performance would be near the minimum level of acceptability and

Table 7

Consistency of Judges Over Two Trials

Rating (Situation)	Trial 1	Trial 2									
		Nurses		Pharmacists		Physicians		Veterinarians		Consumers	
		Not Acceptable	Acceptable	Not Acceptable	Acceptable	Not Acceptable	Acceptable	Not Acceptable	Acceptable	Not Acceptable	Acceptable
1	Acceptable	15.7	60.5	11.8	46.5	11.5	57.7	15.3	35.3	12.3	51.3
	Not Acceptable	17.4	6.5	32.6	9.0	17.9	12.8	34.4	15.0	21.3	15.0
2	Acceptable	15.7	56.8	13.9	45.1	12.2	64.7	16.4	47.2	16.3	43.8
	Not Acceptable	20.4	7.1	32.6	8.3	14.1	9.0	22.8	13.6	27.1	12.8
3	Acceptable	18.5	46.9	16.7	32.5	12.8	52.6	16.1	43.1	14.3	47.0
	Not Acceptable	26.2	8.3	41.7	9.2	20.5	14.1	24.2	16.7	24.7	14.0
4[a]	Acceptable	--	--	--	--	--	--	--	--	17.0	51.0
	Not Acceptable	--	--	--	--	--	--	--	--	19.6	12.5

Note. Tabled entries are percentages of judgments over all judges and all protocols. Ratings (situations) correspond to those listed in Table 1.

[a]Made only for consumers.

Table 8

Inter-Trial Correlations for All Judges

Group	Rating (Situation)		Trial 1	Trial 2	Correlation
Nurses (N = 27)	1	M	3.04	2.82	.55
		SD	.91	.93	
	2	M	2.96	2.77	.60
		SD	.94	1.00	
	3	M	2.79	2.58	.60
		SD	.98	1.00	
Pharmacists (N = 12)	1	M	2.64	2.58	.65
		SD	1.03	1.02	
	2	M	2.68	2.58	.56
		SD	1.03	.98	
	3	M	2.45	2.32	.60
		SD	1.06	1.10	
Physicians (N = 13)	1	M	2.83	2.92	.45
		SD	.93	.96	
	2	M	2.99	3.01	.50
		SD	.92	.97	
	3	M	2.76	2.77	.47
		SD	.91	1.01	
Veterinarians (N = 30)	1	M	2.48	2.47	.40
		SD	.95	.95	
	2	M	2.72	2.73	.45
		SD	.93	.89	
	3	M	2.62	2.68	.41
		SD	.97	.89	
Consumers (N = 51)	1	M	2.77	2.84	.54
		SD	1.09	1.08	
	2	M	2.72	2.59	.50
		SD	1.07	1.13	
	3	M	2.71	2.67	.53
		SD	1.09	1.12	
	4	M	2.87	2.76	.48
		SD	1.05	1.09	

Note. Correlations are based on ratings from 1 to 4 on each trial. They have been computed over all judges and all three tapes.

therefore judgments would be difficult to make. Although these correlations are most certainly too low to make decisions about individual performances, they are sufficiently high to be used in the standard-setting process, in which all examinee ratings were used.

Further evidence of the stability of the standard-setting process is presented below. Relationships between judgments and TSE scores are given for each trial and are presented graphically for visual comparison.

Relationships of Judgments to TSE Scores

Appendix B shows--by rating, trial, and judge group--the percentages of judges that rated each test protocol as acceptable. Table 9 contains the coefficients for the logistic regression functions that were fitted to these data; Figures 1-16 present these functions graphically (and the data points on which they are based). As Table 9 indicates, the logistic regressions of judges' ratings on TSE total scores are, as anticipated, statistically significant in each case; in most cases, the relationships are substantial.

As Figures 1-16 (and Table 9) show, the standard-setting process employed here yielded somewhat different results on each trial, because of the various sources of error mentioned above. However, in the intermediate TSE score ranges few of the analyses revealed substantial between-group differences in the predictions of proportions of judges who rated examinees as acceptable. Comparisons between groups of judges show, however, that the ratings of some groups corresponded more closely with examinees' TSE scores than did those of other groups. These differences may be related to such factors as the differences in the situations that each group was asked to consider and in the particular characteristics and points of view of each group of judges. For each group, however, TSE scores were related significantly to judges' ratings. It is also clear that, in each analysis, some low-scoring examinees were rated as acceptable by relatively large numbers of judges while some high-scoring test takers were considered acceptable by relatively few judges. This lack of perfect correspondence between ratings and test scores is, again, due to several factors. A primary factor is undoubtedly the sampling of only two to three minutes of test performance from each examinee's total of approximately 20 minutes. Thus, the reliable assessment of individual candidates, while an objective of the TSE testing program, was not as important a consideration here.

Table 10 gives the approximate TSE score levels associated with three different standard-setting criteria. The first is the TSE score at the point where judges were most uncertain about the acceptability of speaking performance, i.e., where 50% of the judges were predicted to rate the performance as acceptable. The second and third criteria involved the application of decision-theoretic considerations of the seriousness of each of the two possible types of certification errors. One criterion--

Table 9

Logistic Regression Coefficients for Each Rating for Each Group

Group	Rating	Trial	Coefficients and Standard Errors				
			a	SE_a	b	SE_b	χ^2
Nurses	1	1	.0321	.0123	- 4.43	2.04	10.98***
		2	.0619	.0226	- 9.19	3.62	20.76***
	2	1	.0254	.0105	- 3.52	1.78	8.31**
		2	.0812	.0307	-12.60	4.99	25.35***
	3	1	.0285	.0110	- 4.36	1.92	10.19***
		2	.0571	.0193	-10.60	3.58	21.57***
Pharmacists	1	1	.0243	.0122	- 3.23	2.06	5.54*
		2	.0187	.0104	- 2.96	1.88	4.10*
	2	1	.0338	.0156	- 4.45	2.51	8.01**
		2	.0187	.0104	- 2.96	1.88	4.10**
	3	1	.0106	.0089	- 1.72	1.63	1.56
		2	.0222	.0115	- 4.42	2.21	5.10*
Physicians	1	1	.0473	.0211	- 5.54	3.33	11.28***
		2	.0292	.0137	- 4.02	2.29	7.09**
	2	1	.0585	.0290	- 7.17	4.10	11.30***
		2	.1380	.1330	-16.60	16.80	13.82***
	3	1	.0407	.0177	- 6.20	2.98	10.72***
		2	.0264	.0126	- 3.82	2.17	6.46*
Veterinarians	1	1	.0311	.0116	- 5.76	2.17	11.60***
		2	.0304	.0114	- 5.64	2.14	11.29***
	2	1	.0213	.0095	- 2.99	1.64	6.58**
		2	.0520	.0181	- 8.57	3.12	19.71***
	3	1	.0246	.0101	- 3.86	1.78	8.50**
		2	.0304	.0114	- 5.18	2.06	11.42***
Consumers	1	1	.0256	.0105	- 3.55	1.79	8.41**
		2	.0451	.0163	- 6.51	2.66	15.92***
	2	1	.0311	.0118	- 4.64	2.02	11.18***
		2	.0222	.0095	- 3.59	1.70	7.41**
	3	1	.0557	.0195	- 8.96	3.32	20.54***
		2	.0190	.0090	- 2.60	1.57	5.49*
	4	1	.0288	.0113	- 4.08	1.91	9.88**
		2	.1490	.0710	-24.10	11.70	32.91***

* = $p < .05$
** = $p < .01$
*** = $p < .001$

Note. Logistic regression coefficients a and b are defined as follows: $z = ax + b$ where x is the total TSE score and $y = e^z/(1 + e^z)$ where y is 1 or 0 for pass or fail. A significant χ^2 value indicates that TSE scores made a significant contribution to prediction.

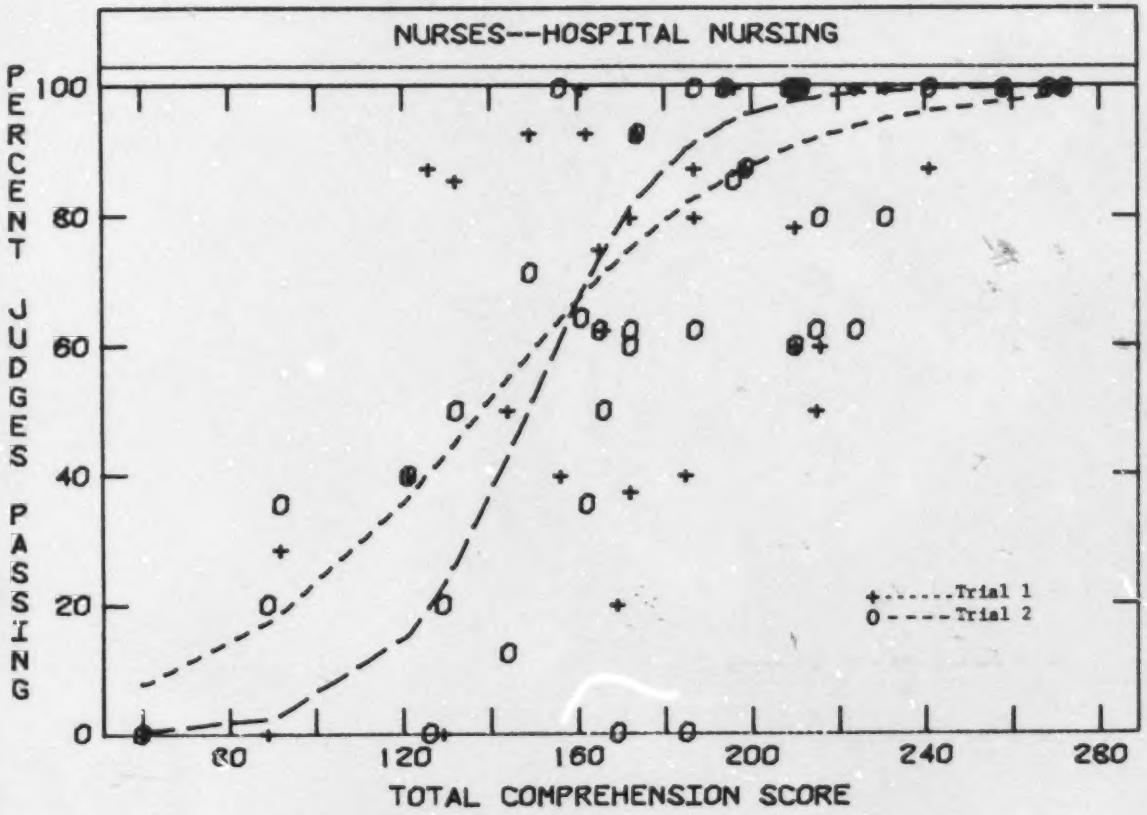

Figure 1. Regression of judges' ratings on TSE total comprehensibility scores.

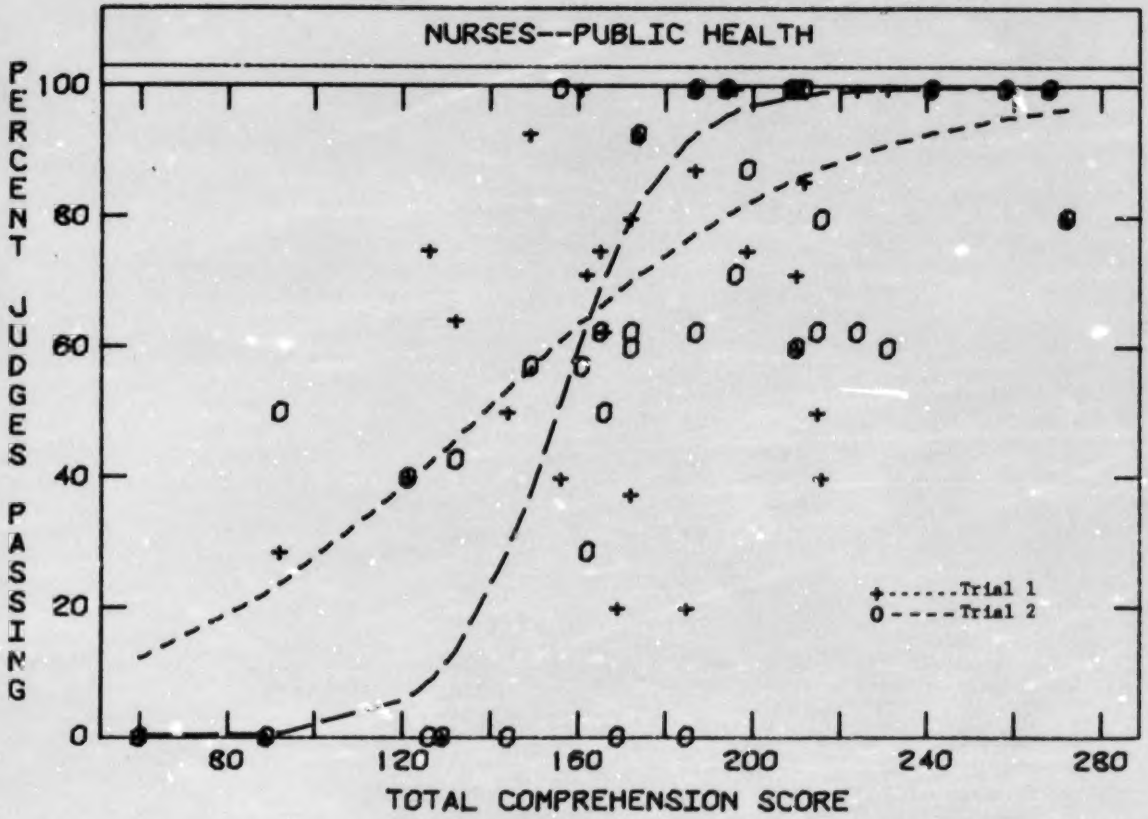

Figure 2. Regression of judges' ratings on TSE total comprehensibility scores.

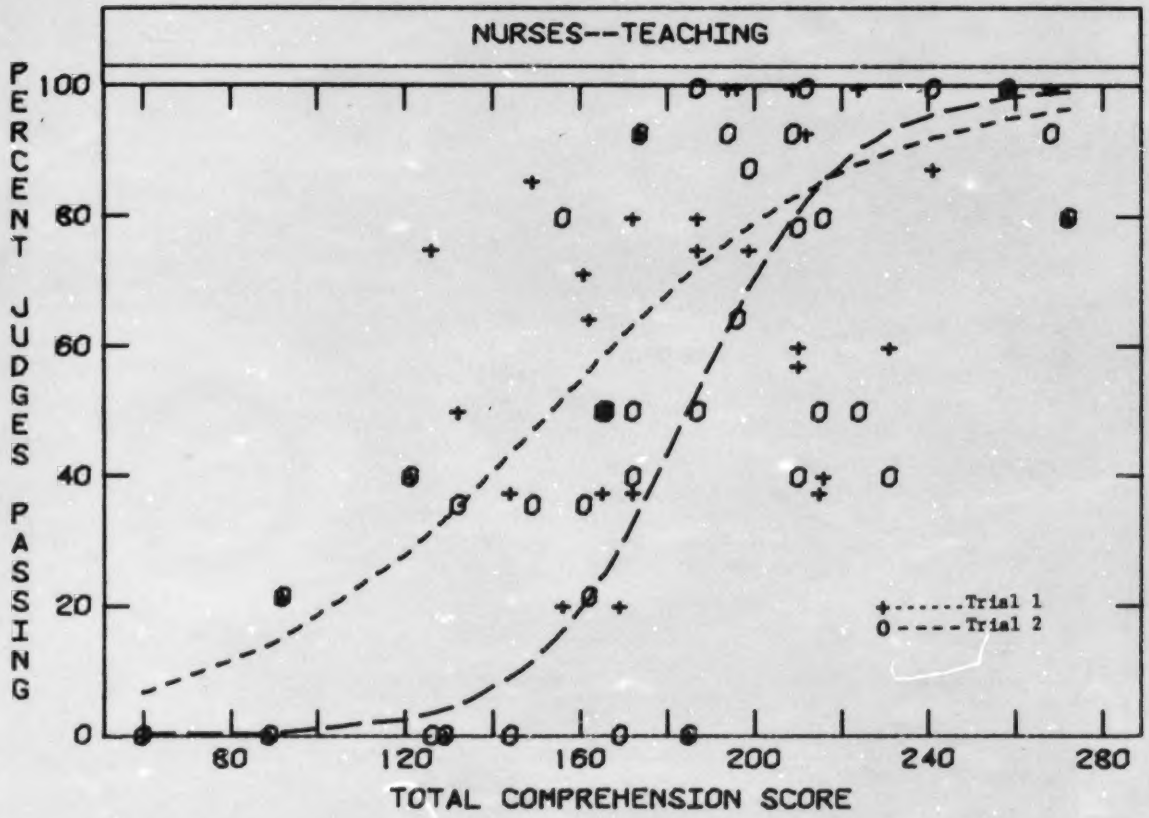

Figure 3. Regression of judges' ratings on TSE total comprehensibility scores.

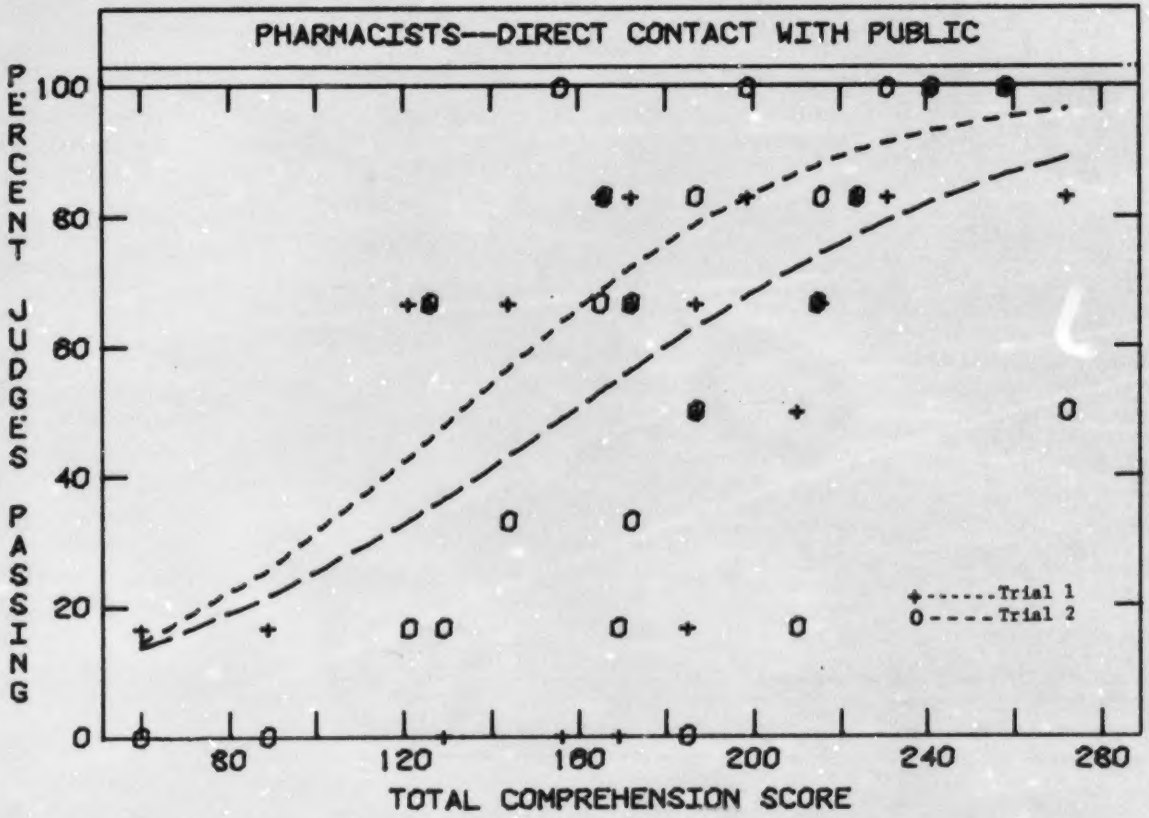

Figure 4. Regression of judges' ratings on TSE total comprehensibility scores.

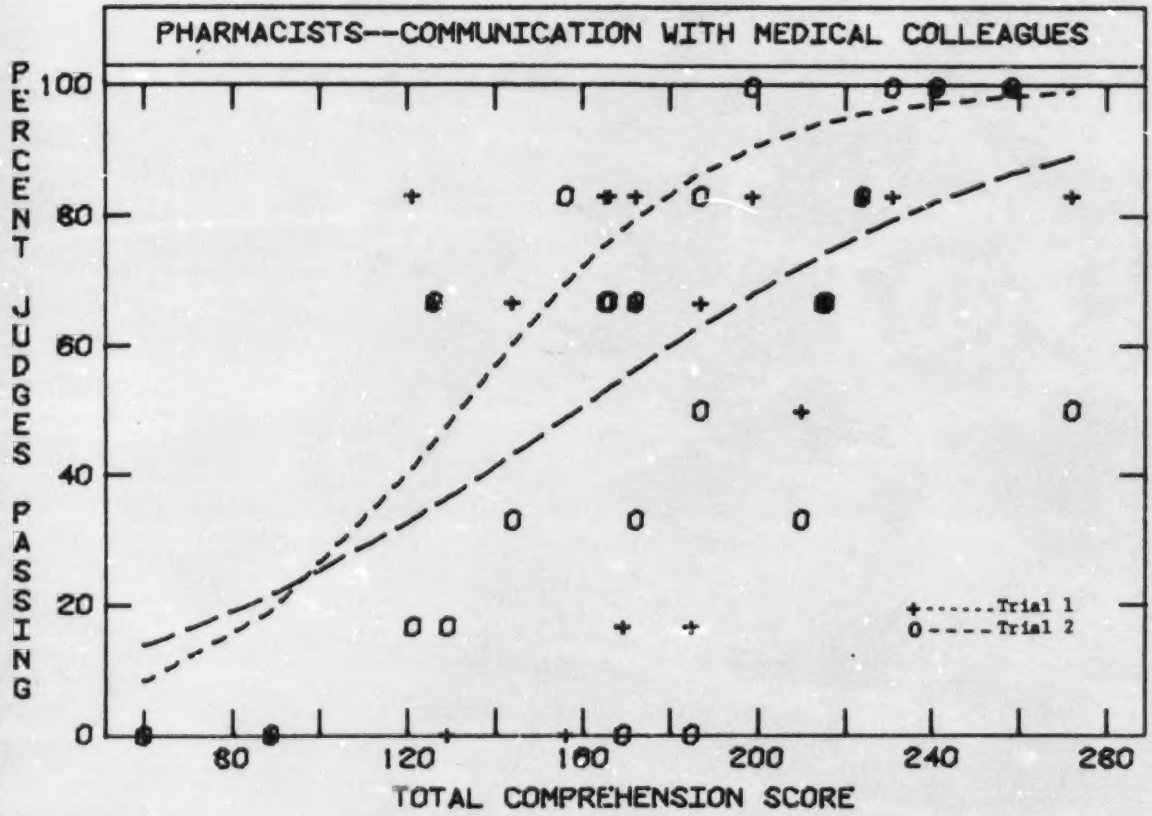

Figure 5. Regression of judges' ratings on TSE total comprehensibility scores.

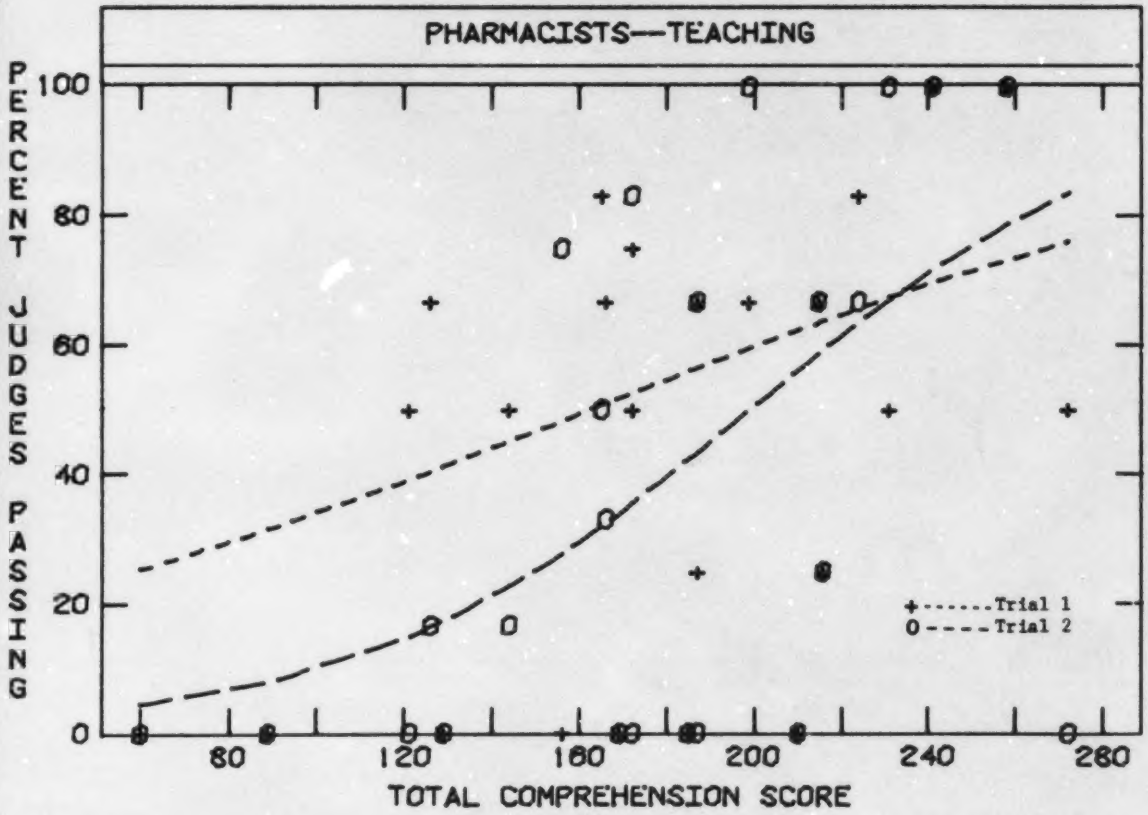

Figure 6. Regression of judges' ratings on TSE total comprehensibility scores.

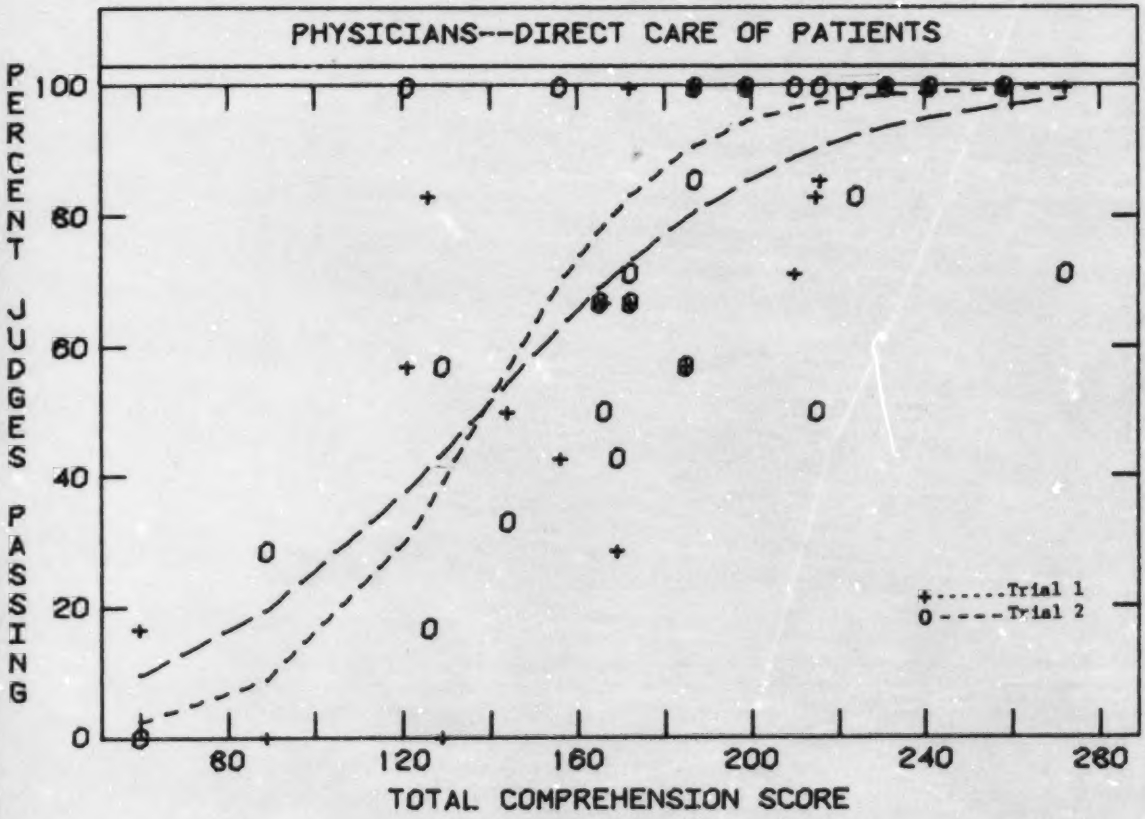

Figure 7. Regression of judges' ratings on TSE total comprehensibility scores.

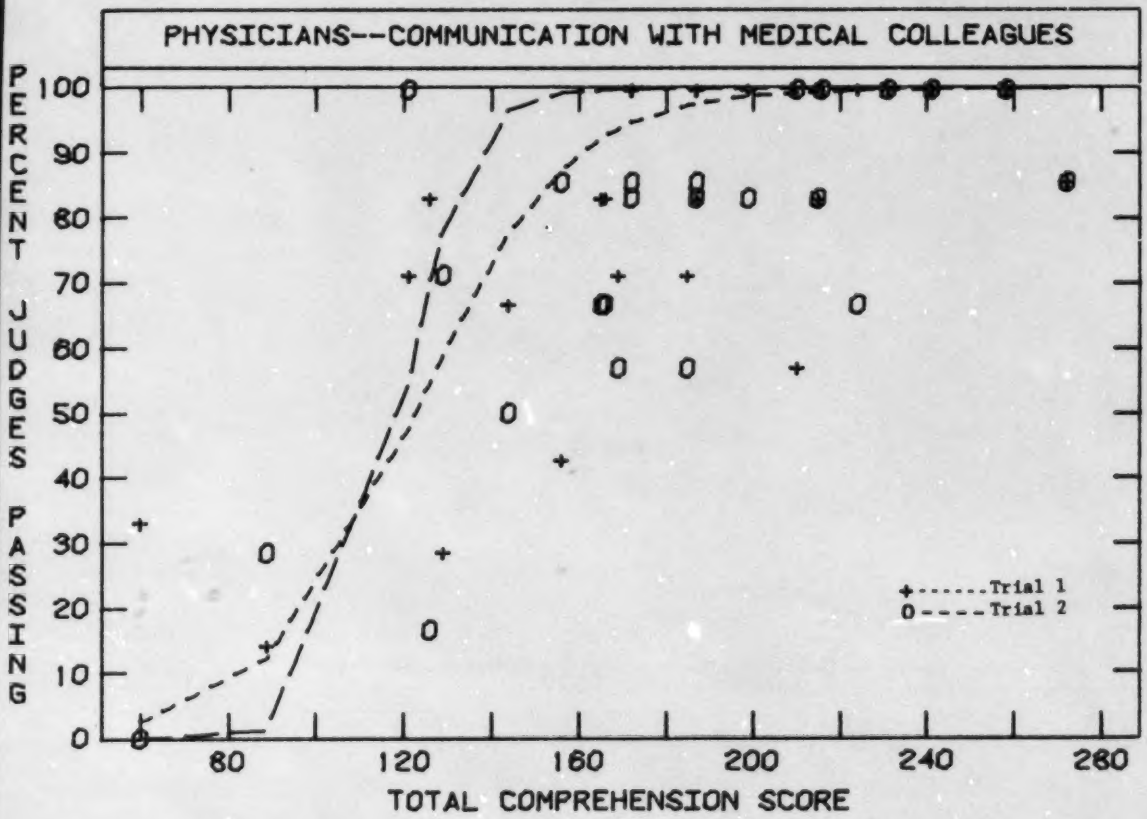

Figure 8. Regression of judges' ratings on TSE total comprehensibility scores.

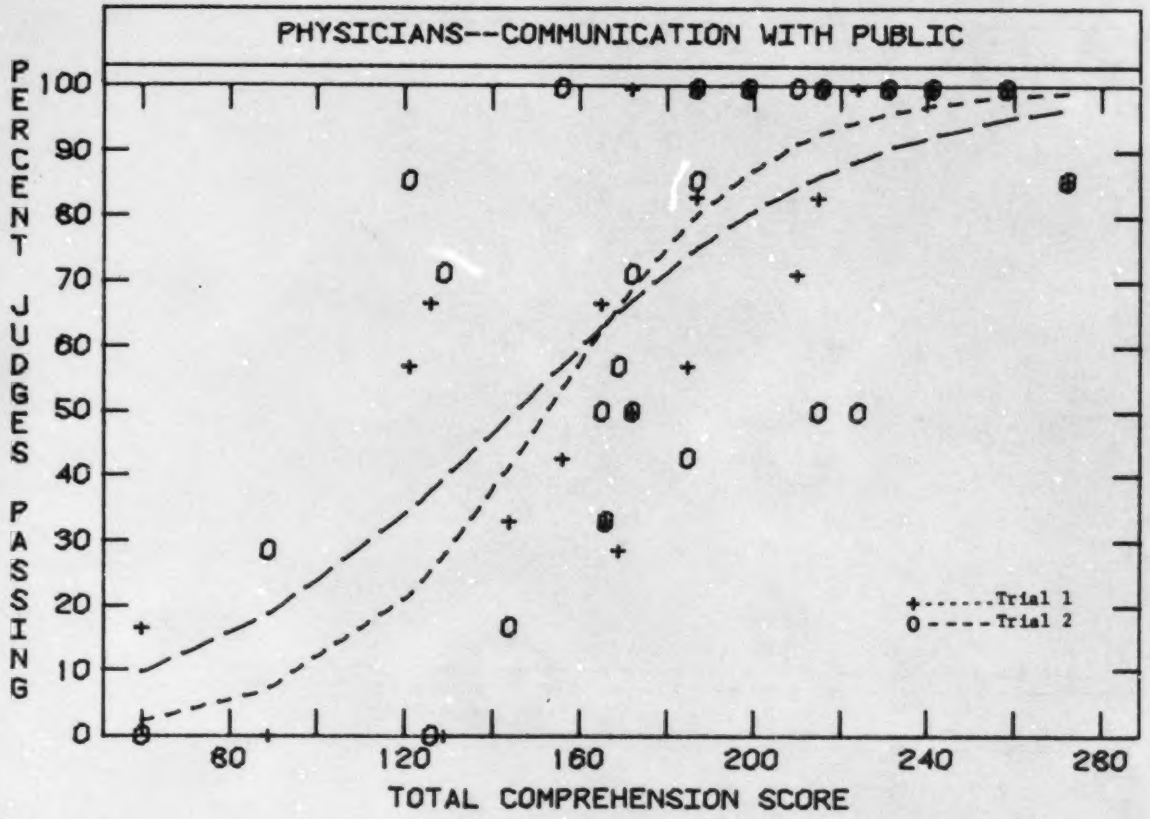

Figure 9. Regression of judges' ratings on TSE total comprehensibility scores.

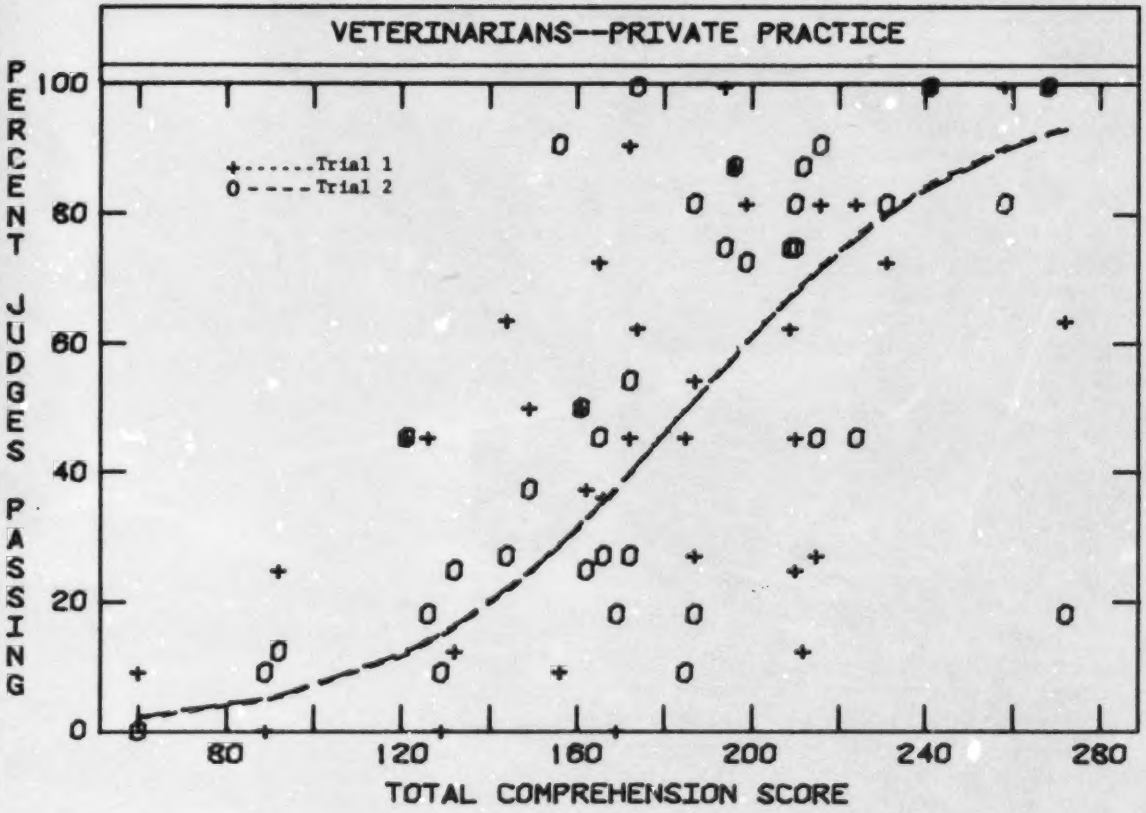

Figure 10. Regression of judges' ratings on TSE total comprehensibility scores.

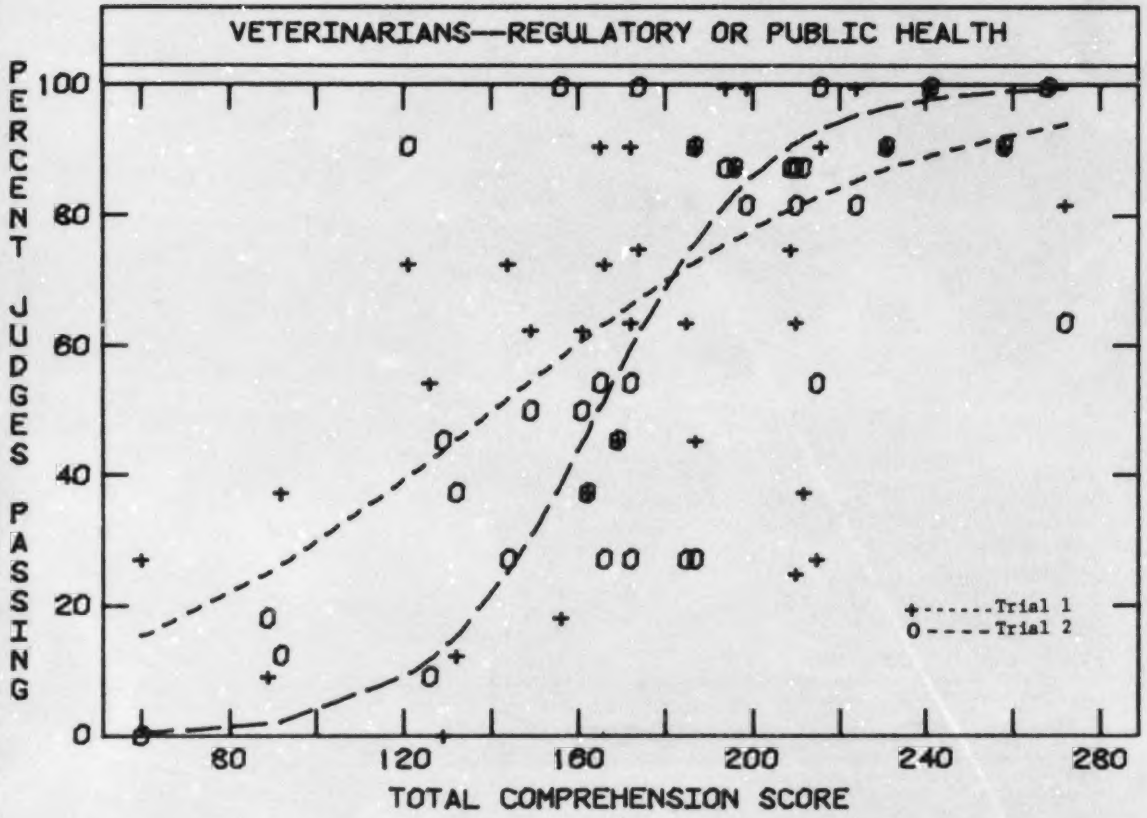

Figure 11. Regression of judges' ratings on TSE total comprehensibility scores.

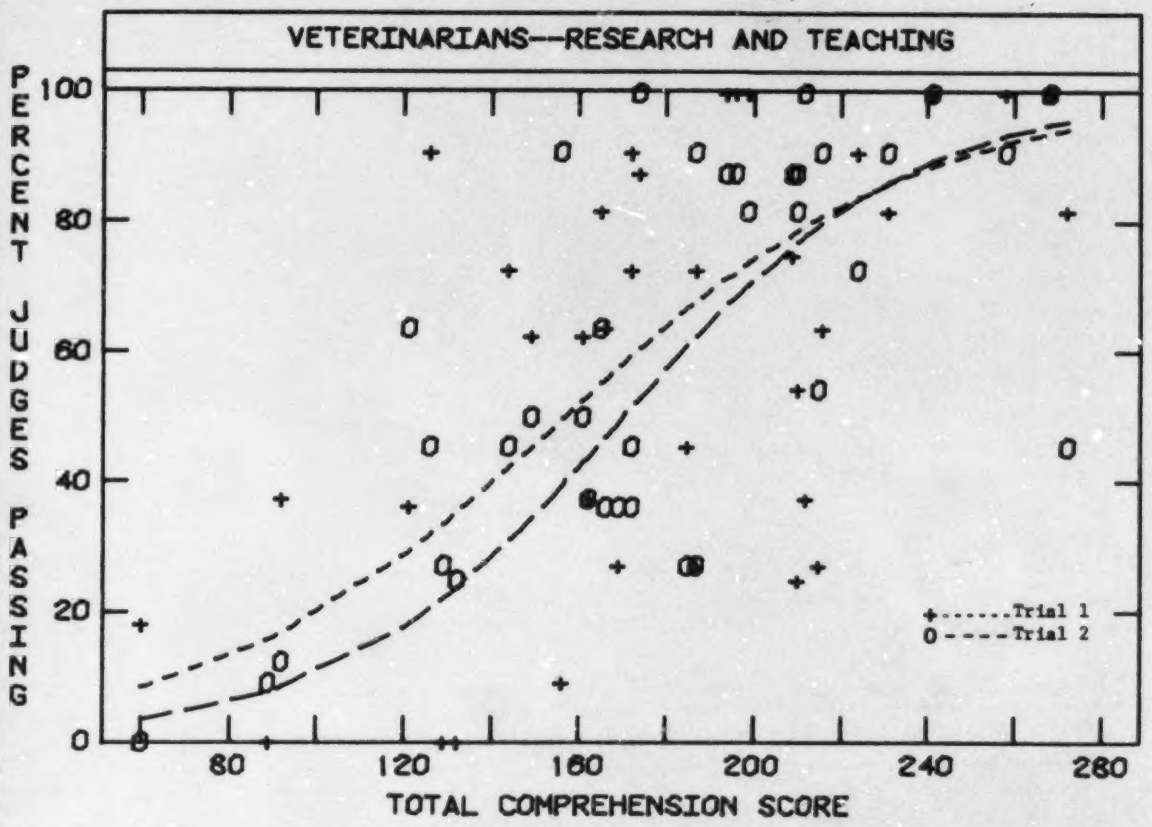

Figure 12. Regression of judges' ratings on TSE total comprehensibility scores.

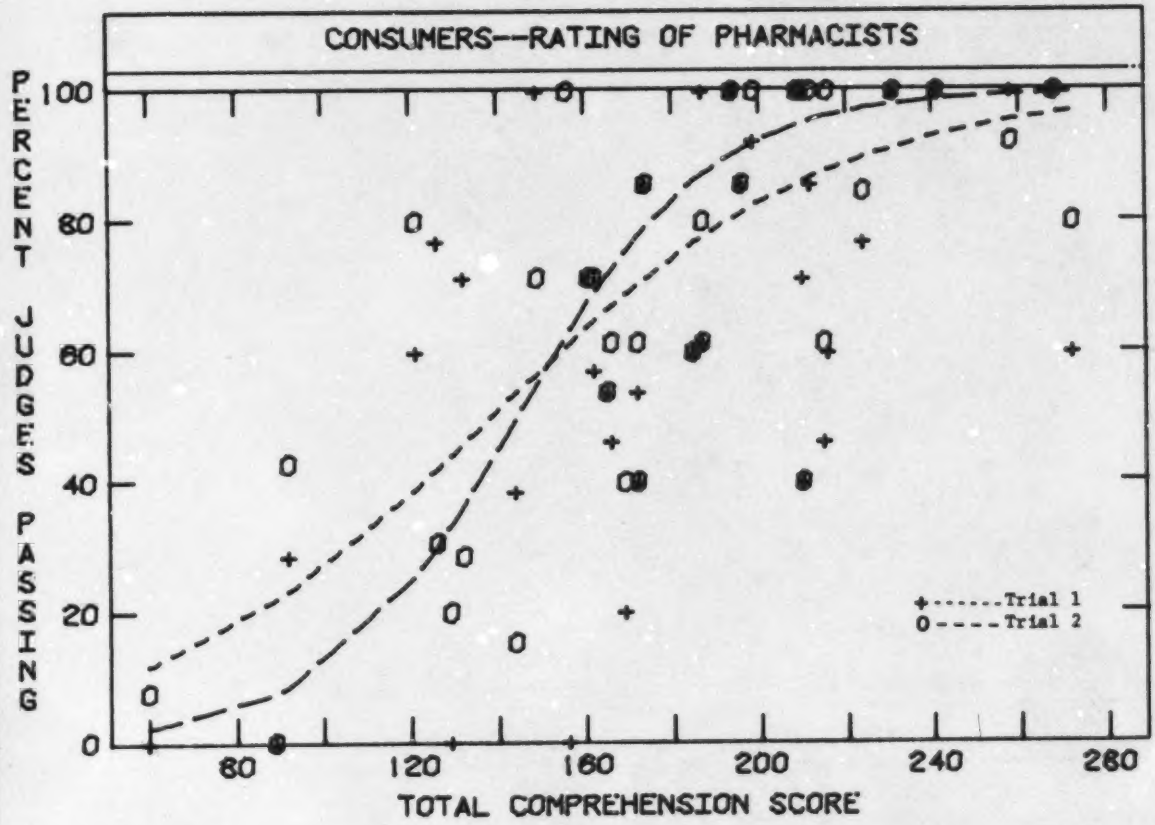

Figure 13. Regression of judges' ratings on TSE total comprehensibility scores.

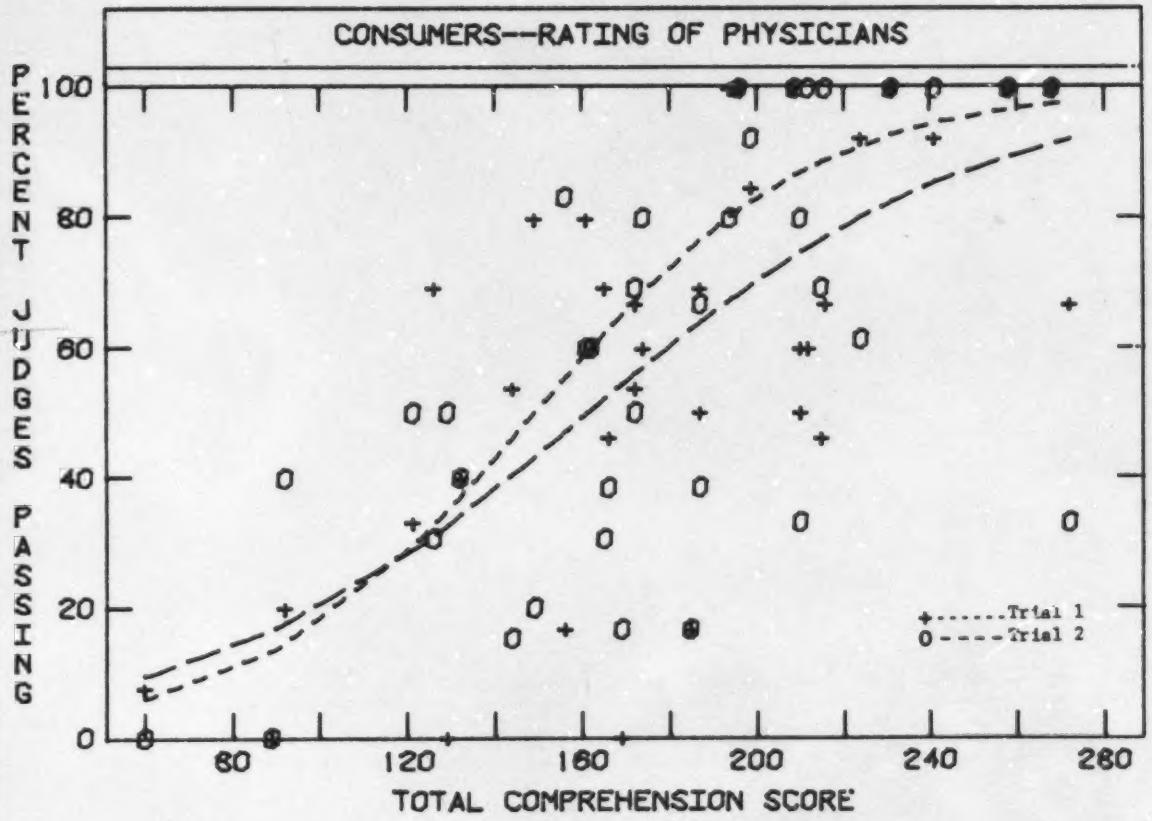

Figure 14. Regression of judges' ratings on TSE total comprehensibility scores.

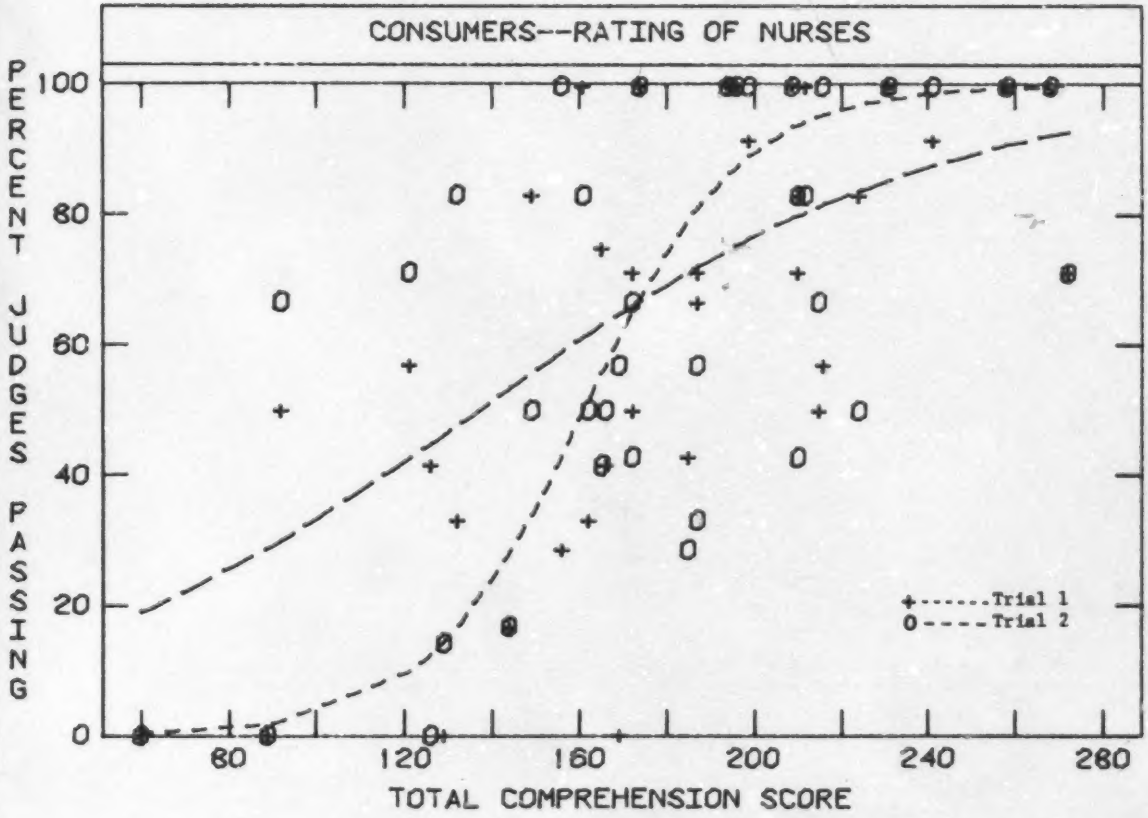

Figure 15. Regression of judges' ratings on TSE total comprehensibility scores.

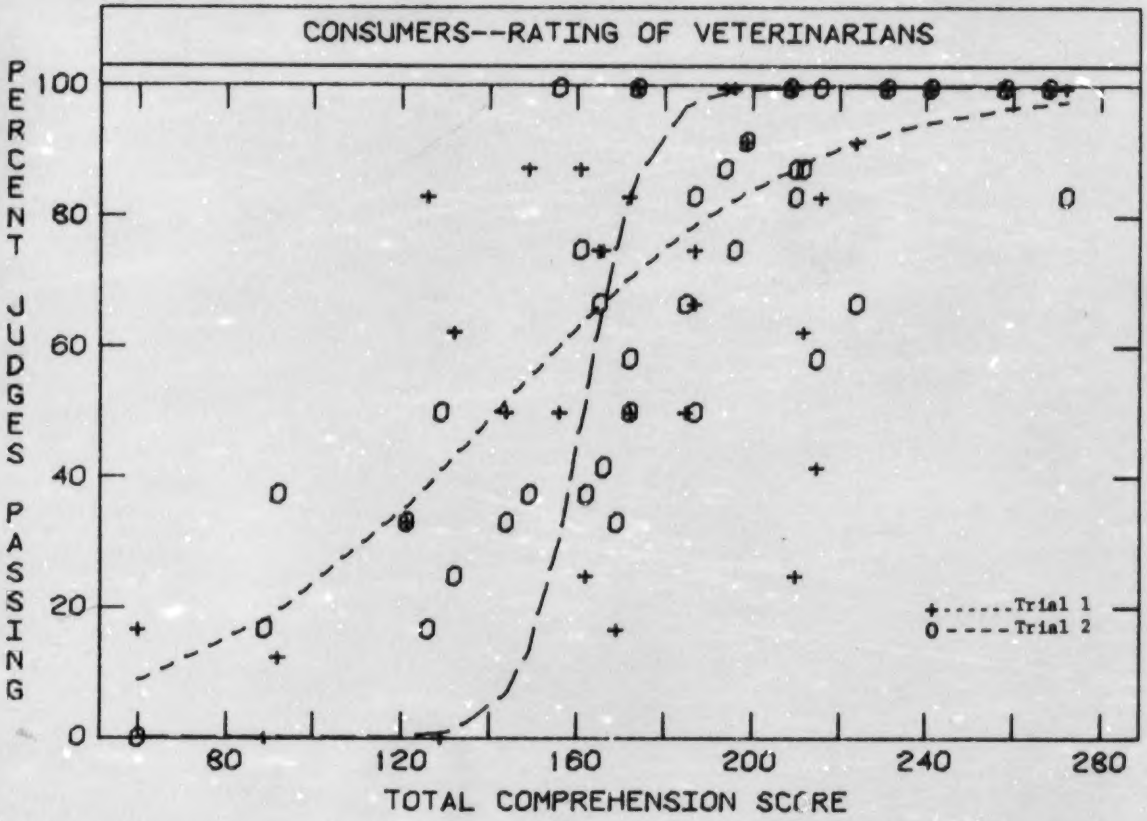

Figure 16. Regression of judges' ratings on TSE total comprehensibility scores.

Table 10

Approximate Standards on the TSE Computed by Three Methods

Group/Rating	Trial	Method: Predicted Probability at 50% Passing	Method: Cost Matrix Methods: Equal Cost	Method: Cost Matrix Methods: Empirical Weights
Nurses				
Hospital nursing	1	140	130 - 145 (.41 - .54)	125 - 130 (.38 - .44)
	2	150	135 - 155 (.28 - .61)	145 - 150 (.44 - .49)
Public health	1	140	130 - 145 (.43 - .53)	130 - 135 (.44 - .46)
	2	155	150 - 155 (.39 - .51)	145 - 150 (.29 - .38)
Teaching	1	155	130 - 160 (.34 - .54)	170 - 175 (.63 - .64)
	2	185	170 - 190 (.34 - .53)	185 - 190 (.51 - .53)
Pharmacists				
Direct contact with public	1	130	*	95 - 125 (.26 - .43)
	2	160	*	150 - 160 (.44 - .48)
Communication with medical colleagues	1	130	*	*
	2	160	*	145 - 160 (.44 - .48)
Teaching	1	165	150 - 180 (.46 - .53)	210 - 240 (.63 - .66)
	2	200	185 - 220 (.43 - .63)	215 - 230 (.58 - .66)
Physicians				
Direct care of patients	1	140	*	155 - 165 (.71 - .78)
	2	140	*	145 - 170 (.56 - .73)
Communication with medical colleagues	1	120	*	90 - 120 (.13 - .48)
	2	120	*	90 - 120 (.03 - .54)
Communication with general public	1	150	*	170 - 185 (.69 - .78)
	2	145	*	165 - 170 (.64 - .66)
Veterinarians				
Private practice	1	185	170 - 195 (.39 - .56)	215 - 225 (.73 - .76)
	2	185	170 - 195 (.36 - .58)	190 - 195 (.53 - .56)
Regulatory medicine or public health	1	140	105 - 140 (.28 - .51)	135 - 140 (.46 - .51)
	2	165	150 - 170 (.31 - .61)	185 - 190 (.78 - .81)
Research and teaching	1	155	125 - 160 (.31 - .54)	130 - 170 (.34 - .59)
	2	170	170 - 200 (.49 - .69)	190 - 195 (.64 - .66)
Consumers				
Pharmacists	1	140	*	Not obtained for consumers
	2	140	130 - 160 (.38 - .63)	Not obtained for consumers
Physicians	1	150	125 - 160 (.31 - .59)	Not obtained for consumers
	2	160	145 - 200 (.41 - .68)	Not obtained for consumers
Nurses	1	160	*	Not obtained for consumers
	2	140	*	Not obtained for consumers
Veterinarians	1	140	*	Not obtained for consumers
	2	160	*	Not obtained for consumers

* Indicates no clear maximum

Note. The predicted probability at 50% passing is score level at which 50% of the judges were predicted to judge the speaker as acceptable. For the equal cost method, the TSE score is the one at which the number of successful predictions (as acceptable or not acceptable) was greatest when the two types of errors are considered equally serious. The empirical weights method uses a cost matrix derived from judges' questionnaire responses. Each type of correct decision and each type of error may have different weights. The TSE score on this method is based on a weighted sum of correct and incorrect decisions. Numbers in parentheses indicate the predicted probability levels for each of the cost matrix methods. All scores have been rounded to intervals of five points.

BEST COPY AVAILABLE

the equal cost criterion[1]--assigned 0 weights to each of the two types of correct decisions and equal (-1) weights to each of the two types of errors. The other cost matrix was derived empirically from judges' responses to a question on the desirability (0 to +5) or undesirability (0 to -5) of each of the four possible licensing or certification outcomes:

(1) licensing a speaker who <u>is</u> at least minimally acceptable

(2) licensing a speaker who <u>is not</u> at least minimally acceptable

(3) <u>not</u> licensing a speaker who <u>is</u> at least minimally acceptable

(4) <u>not</u> licensing a speaker who <u>is not</u> at least minimally acceptable

These assessments were obtained from each group, except consumers, for each rating (Table 11). The score ranges for the latter two criteria were computed by applying these weights to the probabilities of each type of error to arrive at expected utilities. The tabled scores reflect the score levels at which these expected utilities were greatest, i.e., the score levels at which the weighted sums of the more serious errors and the more desirable correct decisions were maximized.

Table 11 shows variation both within and between groups of judges with respect to their perceptions of the tolerability of each possible outcome. In general, each group showed more concern that the correct decision be made for unacceptable speakers: licensing unacceptable speakers tended to be rated as the more serious error and not licensing an unacceptable speaker as the more tolerable of the correct decisions. The sole exception was physicians' ratings of communication with medical colleagues: not licensing an acceptable speaker in this particular situation was viewed as slightly more serious than licensing an unacceptable speaker. Physicians also tended to consider not licensing an acceptable speaker as more serious than did any of the other groups of judges.

Table 10 gives the ranges of TSE scores for each of the three methods of obtaining standards. (It should be noted that readers could select their own probability of judges passing and compute the corresponding levels of TSE scores from Figures 1-16. Generally, the more seriously the misclassification of unacceptable speakers is regarded, the higher the percent of judges passing should be set, and therefore the higher the TSE level.) In some instances the expected utility functions suggested no clear range of scores in which the weighted sum of more serious errors and desirable outcomes would be at a maximum.

---

[1] As Livingston (1982) has pointed out to us, under the equal cost criterion the test score cutoff that minimizes the total cost is algebraically equivalent to the score at which the predicted probability of judges passing is 50% (method 1). Thus, methods 1 and 2 are in a sense redundant. We have included the equal cost results primarily to show the ranges of scores in which the costs are "nearly minimized."

Table 11

Cost Matrices Derived for Each Group of Judges

Group/Situation	Judgment of Speaking Proficiency	Licensure Decision: Yes M	Yes SD	No M	No SD
Nurses					
Hospital nursing	Acceptable	3.33	1.68	-2.69	1.81
	Not acceptable	-4.59	1.10	4.48	1.26
Public health	Acceptable	3.07	1.65	-2.52	1.66
	Not acceptable	-4.67	1.05	4.52	1.23
Teaching	Acceptable	2.74	1.84	-2.19	1.70
	Not acceptable	-4.70	1.05	4.56	1.23
Pharmacists					
Direct contact with public	Acceptable	1.58	1.80	-1.83	1.86
	Not acceptable	-2.58	1.93	1.92	2.10
Communication with medical colleagues	Acceptable	1.83	1.62	-1.75	1.79
	Not acceptable	-2.58	1.93	2.00	2.12
Teaching	Acceptable	1.00	1.00	-1.58	1.75
	Not acceptable	-3.50	2.14	2.42	2.18
Physicians					
Direct care of patients	Acceptable	2.69	1.59	-4.00	1.24
	Not acceptable	-4.38	1.39	3.92	1.80
Communication with medical colleagues	Acceptable	2.83	1.57	-4.08	1.49
	Not acceptable	-3.31	1.81	2.69	1.98
Communication with general public	Acceptable	2.25	1.48	-3.31	1.54
	Not acceptable	-3.62	2.06	3.15	2.25
Veterinarians					
Private practice	Acceptable	2.32	1.83	-2.57	1.76
	Not acceptable	-4.07	1.53	3.82	1.89
Regulatory medicine or public health	Acceptable	2.46	1.40	-2.57	1.70
	Not acceptable	-3.64	1.14	3.39	1.86
Research and teaching	Acceptable	2.55	1.71	-2.48	1.77
	Not acceptable	-3.45	1.67	3.45	1.92

Although Table 10 suggests that estimates of minimum proficiency on the TSE fall in a relatively wide range, many of those at the 50% predicted probability level fall in the 140-160 range. Consumers' ratings of each of the four health professions were remarkably consistent around this level.

For the two professions, nursing and pharmacy, in which teaching was rated, the levels suggested by the analyses were slightly higher (155-200). Veterinarians rated a situation combining "teaching and research," instead of teaching only, and their ratings (155-170) also fell in this range, but at the lower end. Physicians tended to rate the level of oral proficiency required in communication with their medical colleagues as slightly lower than any situation rated by any other group. It should be reiterated that the levels discussed here constitute our best estimation of the minimum levels of oral proficiency required in each situation. They do not, however, necessarily suggest appropriate cutoff scores, which must be determined in light of the consequences of each type of decision error. That the correct classification of not acceptable speakers was viewed as more important than the correct classification of acceptable speakers suggests that the appropriate cutoff scores should lie somewhat above the 50% probability level. For example, to decrease the likelihood that unacceptable speakers would be certified on the basis of TSE results, the TSE cutoff should be higher than the scores corresponding to minimum levels of proficiency (Table 10).

Based on the impressions of TSE test development staff, verbal descriptions have previously been associated with four broad TSE score intervals and reported to examinees (Educational Testing Service, 1980). The intervals and descriptions are as follows:

0 to 49	Overall comprehensibility too low in even the simplest type of speech.
50 to 149	Generally not comprehensible due to frequent pauses and/or rephrasing, pronunciation errors, limited grasp of vocabulary, and lack of grammatical control.
150 to 249	Comprehensible, with a limited number of errors in pronunciation, grammar, choice of vocabulary items, or infrequent pauses or rephrasing.
250 to 300	Completely comprehensible in normal speech, with occasional grammaticial or pronunciation errors in very colloquial phrases.

The TSE score that divides the two "not comprehensible" intervals (0-49 and 50-149) from the two "comprehensible" categories (150-249 and 250-300) is 150, which corresponds closely with the score levels suggested by the analyses presented above, thus providing further empirical support to these verbal descriptions.

Summary and Discussion

The ability to communicate effectively is considered critical to successful performance in numerous occupations, including many health-related professions. One aspect of communicative competence, oral proficiency, is necessary in some degree for performing such tasks as interacting with medical colleagues and counseling or instructing patients. The importance of proficiency in oral communication is evidenced by a growing number of states' and certification agencies' requirements for minimum proficiency in a variety of occupations.

The increasing interest in oral proficiency during the past decade prompted Educational Testing Service to undertake the development of the Test of Spoken English, a standardized test of speaking proficiency of nonnative speakers of English. The test has been validated for the selection of nonnative teaching assistants applying to U.S. institutions.

The objectives of the study reported here were to

(a) provide additional validation of the TSE for the selection and certification of nonnative health professionals

(b) establish a set of procedures for determining standards of proficiency on the TSE

(c) try out these procedures in selected health professions in which the test could be used

Groups of practitioners in four health professions--nursing, pharmacy, medicine, and veterinary medicine--judged the acceptability of the speaking performance of examinees who had taken the TSE. A sample of examinees was chosen to represent various TSE score levels and language groups. Each group of judges was asked to provide for their profession ratings for three distinct situations, such as communicating with professional colleagues. Judges were to indicate whether, in their view, each examinee was at least minimally acceptable to function in each of the situations. Judges' ratings were then related to the TSE scores that had been assigned previously to each examinee's test performance. A larger group of medical care consumers gave more global ratings to the same examinees for each of the four professions.

Background data suggested that the groups of judges were heterogeneous with regard to a number of relevant personal characteristics and experiences, such as degree of exposure to nonnative speakers. This diversity suggested that a variety of points of view were probably represented by these judges.

Analyses of judges' ratings revealed a substantial degree of consistency in judges' responses. Even though only relatively small segments of examinees' test performances were rated, judges tended to assign the same ratings when asked to make a second set of ratings of a second sample of

test performances for the same examinees. Likewise, the relationships of judges' ratings to TSE scores generally exhibited considerable stability, although the data clearly showed that the procedures applied here can sometimes produce substantially different results from one occasion to the next.

Alternative methods were applied to determine possible ranges of cutoff scores on the TSE. One of these methods involved a consideration of the consequences of decision outcomes, e.g., whether the certification or licensing of an unacceptable speaker is more or less serious than failing to certify an acceptable speaker. A range of possible cutoff scores was computed for each situation within each profession.

Other Considerations

A number of social and political issues surround the development and implementation of standards. Among the issues considered by Livingston and Zieky (1982) are the following:

Allowing exceptions to the decision rule. The use of a passing score that does not allow for exceptions may be too rigid, or at least seem so. For the TSE, as for other measures, examinees having comparable levels of skill or ability (in this case, speaking proficiency) may not get comparable test scores. If, as Livingston and Zieky point out, one could anticipate all the possible reasons that would justify exceptions to passing scores, these reasons could become part of the decision rule. Unfortunately, such extraordinary foresight is extremely rare. However, established procedures (e.g., a standing committee to review requests for exceptions) could be developed to deal with such events.

Retesting speakers. Because speaking proficiency can be expected to improve with both practice and specific instruction (e.g., Hinofotis & Bailey, 1981; Pederson, 1979), TSE examinees who do not meet minimum standards should probably be permitted to take the test again. They should also be allowed to retake the test because of the possibility of merely having had a "bad day." Because a relatively limited number of questions and alternative forms are available, the number of retestings may need to be restricted. Such restriction might serve as an incentive for examinees to practice, to seek formal instruction, or to prepare in other ways for the test more earnestly than they would if unlimited retesting were available.

Using normative data to set standards. As with most standard-setting issues, no clear-cut answer is possible here. Some would argue that norms should have no role in determining standards. Normative data have not entered into in the standard-setting project described here. However, such data may serve as a useful "reality check" once a standard is put into practice. Knowing that nearly all test takers exceed a particular standard might suggest raising the standard, and knowing that very few examines pass the cutoff could suggest lowering it.

Allowing changes in standards over time. Some circumstances may suggest the desirability of raising or lowering standards over time. Changes in the nature of a profession and in the amount and quality of the verbal interaction it requires might suggest the need to revise a standard, as might shortages of trained personnel in a particular field. Thus, no standard should be considered fixed forever.

Setting different standards for different groups. Because different levels of speaking proficiency may be required of particular specialties within a medical profession, setting different standards for individual specialties or sets of specialties might be reasonable, although the feasibility of implementing such a system would undoubtedly be considerably less. The ranges of standards implied by the analyses conducted in this study suggest a certain degree of generalizability across job situations; that is, in general, the standards do not differ dramatically for the various situations within a particular profession, at least for the general situations considered in this study. Although this finding would seem to mitigate against different standards, there may be some specific situations for which a higher or lower standard is appropriate.

Next Steps

In most instances the analyses presented here show a strong relationship between TSE scores and judges' ratings of the acceptability of speakers. Although no clear rules exist, this relationship appears strong enough to permit setting standards of proficiency on the test. The results presented here give a relatively good indication of the ranges of TSE scores that might be considered acceptable in several health-related professions. There are, however, a number of reasons why certification agencies and other TSE users may choose to conduct their own studies. One is simply to verify our results. Other, more cogent reasons are to base the standards on more appropriate groups of judges than used here (e.g., judges more representative of the national membership of a particular organization). Also, some agencies may wish to judge the adequacy of performance for situations different from the ones we have specified.

In summary, this study has provided additional evidence for the validity of the Test of Spoken English as a measure of oral proficiency and has made some degree of progress toward setting appropriate standards of proficiency on the test. The study has not, however, resulted in any clearly specified cutoff scores, in recognition that "all standard-setting methods are arbitrary and fallible" (Shepard, 1980). Although the procedures applied here have moved potential TSE users closer to rational and defensible cutoffs that would meet commonly accepted professional standards for test use (American Psychological Association, 1974), in the final analysis such cutoff scores can be established only by authorized certification agencies. Nevertheless, the procedures and results presented above should reduce considerably the inevitable arbitrariness in setting standards of proficiency on the Test of Spoken English.

References

American Psychological Association. Standards for educational and psychological tests. Washington, D.C.: Author, 1974.

Carroll, J. B.; Clark, J. L. D.; Edwards, T.; & Handrick, F. The foreign language attainments of language majors in the senior year: A survey conducted in U.S. colleges and universities. Cambridge, MA: Harvard University, 1967.

Clark, J. L. D., & Swinton, S. S. The Test of Spoken English as a measure of communicative ability in English-medium instructional settings (TOEFL Research Report 7). Princeton, NJ: Educational Testing Service, 1980.

Clark, J. L. D., & Swinton, S. S. An exploration of speaking proficiency measures in the TOEFL context (TOEFL Research Report 4). Princeton, NJ: Educational Testing Service, 1979.

Educational Testing Service. TSE 1981-82 Information Bulletin. Princeton, NJ: Author, 1980.

Hinofotis, F. B., & Bailey, K. M. American undergraduates' reactions to the communication skills of foreign teaching assistants. In J. C. Fisher, M. A. Clark, & J. Schachter (Eds.), On TESOL' 80. Washington, DC: Teachers of English to Speakers of Other Languages, 1981.

Livingston, S. A. Setting standards of speaking proficiency. In J. L. D. Clark (Ed.), Direct Testing of Speaking Proficiency: Theory and Application. Princeton, NJ: Educational Testing Service, 1978.

Livingston, S. A. Personal communication, September 7, 1982.

Livingston, S. A., & Zieky, M. J. Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service, 1982.

Osanyinbi, J. A concurrent validity study of the West African School Certificate and General Certificate of Education English Language Examination. Unpublished doctoral dissertation, University of Wisconsin, 1974.

Pederson, L. M. Foreign medical graduates: Identification of specific problems and methods for improving communication skills. Paper presented at the Speech Communication Association Summer Conference on Intercultural Communication, Honolulu, HI, August 1979. (ED 180 035)

Pike, L. W. An evaluation of alternative item formats for testing English as a foreign language (TOEFL Research Report 2). Princeton, NJ: Educational Testing Service, 1979.

Schumann, J. Second language acquisition research: Getting a more global look at the learner. In H. D. Brown (Ed.), Papers in Second Language Acquisition Language Learning. Special Issue No. 4, January, 1976, 15-28.

Shephard, L. Technical issues in minimum competency testing. In D. C. Berliner (Ed.), Review of Research in Education, Vol. 8. Itasca, IL: F. E. Peacock Publishers, Inc., 1980.

Showers, B., & Nyenhuis, W. Job simulation test of English language proficiency for foreign trained nurses. Paper presented at the annual conference of the International Personnel Managers Association Assessment Council, Minneapolis, June, 1982.

Sollenberger, H. E. Development and current use of the FSI oral interview test. In J. L. D. Clark (Ed.), Direct Testing of Speaking Proficiency: Theory and Application. Princeton, NJ: Educational Testing Service, 1978.

Appendix A

Arrangement of TSE Standard Setting Tapes

Tape 1		Tape 2		Tape 3	
Examinee	Section	Examinee	Section	Examinee	Section
A	2	A	2	A	2
B	6	B	6	B	6
C	7	C	7	C	7
1	6	13	6	25	2
2	2	14	6	26	6
3	2	15	7	27	2
4	7	16	7	28	7
5	6	17	7	29	2
6	7	18	2	30	6
7	2	19	7	31	2
8	2	20	6	32	7
9	7	21	6	33	7
10	6	22	2	34	6
11	7	23	2	35	6
12	6	24	2	36	7
3	7	18	6	35	2
2	7	21	7	26	7
9	2	14	2	31	7
5	2	16	6	29	7
10	7	19	2	25	6
8	6	24	6	32	6
12	7	17	6	36	6
7	6	23	7	33	2
1	2	20	2	30	7
11	6	22	7	27	6
6	6	15	2	34	2
4	2	13	7	28	2

3 sets of 12 examinees with scores distributed as follows on each tape.

TSE Score Level	Frequency
0 - 49	0
50 - 99	1
100 - 149	2
150 - 174	3
175 - 199	2
200 - 224	2
225 - 249	1
250 - 300	1

Note. Examinees A, B, and C are unanalyzed practice protocols.

Appendix B

Percentages of Judges Rating Each Speaker as Acceptable by TSE Score Level

		Nurses		Pharmacists		Physicians		Veterinarians		Consumers	
TSE Score	Rating	Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2
60	1	0.0	0.0	16.7	0.0	16.7	0.0	9.1	0.0	0.0	7.7
	2	0.0	0.0	0.0	0.0	33.3	0.0	27.3	0.0	7.7	0.0
	3	0.0	0.0	0.0	0.0	16.7	0.0	18.2	0.0	0.0	0.0
	4	--	--	--	--	--	--	--	--	16.7	0.0
89	1	0.0	20.0	16.7	0.0	0.0	28.6	0.0	9.1	0.0	0.0
	2	0.0	0.0	0.0	0.0	14.3	28.6	9.1	18.2	0.0	0.0
	3	0.0	0.0	0.0	0.0	0.0	28.6	0.0	9.1	0.0	0.0
	4	--	--	--	--	--	--	--	--	0.0	16.7
92	1	28.6	35.7					25.0	12.5	28.6	42.9
	2	28.6	50.0					37.5	12.5	20.0	40.0
	3	21.4	21.4					37.5	12.5	50.0	66.7
	4	--	--					--	--	12.5	37.5
121	1	40.0	40.0	66.7	16.7	57.1	100.0	45.5	45.5	60.0	80.0
	2	40.0	40.0	83.3	16.7	71.4	100.0	72.7	90.9	33.3	50.0
	3	40.0	40.0	50.0	0.0	57.1	85.7	36.4	63.6	57.1	71.4
	4	--	--	--	--	--	--	--	--	33.3	33.3
126	1	87.5	0.0	66.7	66.7	83.3	16.7	45.5	18.2	76.9	30.8
	2	75.0	0.0	66.7	66.7	83.3	16.7	54.5	9.1	69.2	30.8
	3	75.0	0.0	66.7	16.7	66.7	0.0	90.9	45.5	41.7	0.0
	4	--	--	--	--	--	--	--	--	83.3	16.7
129	1	0.0	20.0	0.0	16.7	0.0	57.1	0.0	9.1	0.0	20.0
	2	0.0	0.0	0.0	16.7	28.6	71.4	0.0	45.5	0.0	50.0
	3	0.0	0.0	0.0	0.0	0.0	71.4	0.0	27.3	0.0	14.3
	4	--	--	--	--	--	--	--	--	0.0	50.0
132	1	85.7	50.0					12.5	25.0	71.4	28.6
	2	64.3	42.9					12.5	37.5	40.0	40.0
	3	50.0	35.7					0.0	25.0	33.3	83.3
	4	--	--					--	--	62.5	25.0
144	1	50.0	12.5	66.7	33.3	50.0	33.3	63.6	27.3	38.5	15.4
	2	50.0	0.0	66.7	33.3	66.7	50.0	72.7	27.3	53.8	15.4
	3	37.5	0.0	50.0	16.7	33.3	16.7	72.7	45.5	16.7	16.7
	4	--	--	--	--	--	--	--	--	50.0	33.3
149	1	92.9	71.4					50.0	37.5	100.0	71.4
	2	92.9	57.1					62.5	50.0	80.0	20.0
	3	85.7	35.7					62.5	50.0	83.3	50.0
	4	--	--					--	--	87.5	37.5

TSE Score	Rating	Nurses		Pharmacists		Physicians		Veterinarians		Consumers	
		Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2
156	1	40.0	100.0	0.0	100.0	42.9	100.0	9.1	90.9	0.0	100.0
	2	40.0	100.0	0.0	83.3	42.9	85.7	18.2	100.0	16.7	83.3
	3	20.0	80.0	0.0	75.0	42.9	100.0	9.1	90.9	28.6	100.0
	4	--	--	--	--	--	--	--	--	50.0	100.0
161	1	100.0	64.3					50.0	50.0	71.4	71.4
	2	100.0	57.1					62.5	50.0	80.0	60.0
	3	71.4	35.7					62.5	50.0	100.0	83.3
	4	--	--					--	--	87.5	75.0
162	1	92.9	35.7					37.5	25.0	57.1	71.4
	2	71.4	28.6					37.5	37.5	60.0	60.0
	3	64.3	21.4					37.5	37.5	33.3	50.0
	4	--	--					--	--	25.0	37.5
165	1	75.0	62.5	83.3	66.7	66.7	66.7	72.7	45.5	53.8	53.8
	2	75.0	62.5	83.3	66.7	83.3	66.7	90.9	54.5	69.2	30.8
	3	37.5	50.0	83.3	50.0	66.7	50.0	81.8	63.6	75.0	41.7
	4	--	--	--	--	--	--	--	--	75.0	66.7
166	1	62.5	50.0	83.3	83.3	66.7	50.0	36.4	27.3	46.2	61.5
	2	62.5	50.0	83.3	66.7	83.3	66.7	72.7	27.3	46.2	38.5
	3	50.0	50.0	66.7	33.3	33.3	33.3	63.6	36.4	41.7	50.0
	4	--	--	--	--	--	--	--	--	75.0	41.7
169	1	20.0	0.0	0.0	16.7	28.6	42.9	0.0	18.2	20.0	40.0
	2	20.0	0.0	16.7	0.0	71.4	57.1	45.5	45.5	0.0	16.7
	3	20.0	0.0	0.0	0.0	28.6	57.1	27.3	36.4	0.0	57.1
	4	--	--	--	--	--	--	--	--	16.7	33.3
172	1	37.5	62.5	66.7	66.7	66.7	66.7	45.5	27.3	53.8	61.5
	2	37.5	62.5	66.7	66.7	100.0	83.3	63.6	27.3	53.8	69.2
	3	37.5	50.0	50.0	83.3	50.0	50.0	72.7	45.5	50.0	66.7
	4	--	--	--	--	--	--	--	--	50.0	58.3
172	1	80.0	60.0	83.3	33.3	100.0	71.4	90.9	54.5	40.0	40.0
	2	80.0	60.0	83.3	33.3	100.0	85.7	90.9	54.5	66.7	50.0
	3	80.0	40.0	75.0	0.0	100.0	71.4	90.9	36.4	71.4	42.9
	4	--	--	--	--	--	--	--	--	83.3	50.0
174	1	92.9	92.9					62.5	100.0	85.7	85.7
	2	92.9	92.9					75.0	100.0	60.0	80.0
	3	92.9	92.9					87.5	100.0	100.0	100.0
	4	--	--					--	--	100.0	100.0

TSE Score	Rating	Nurses		Pharmacists		Physicians		Veterinarians		Consumers	
		Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2
185	1	40.0	0.0	16.7	0.0	57.1	57.1	45.5	9.1	60.0	60.0
	2	20.0	0.0	16.7	0.0	71.4	57.1	63.6	27.3	16.7	16.7
	3	0.0	0.0	0.0	0.0	57.1	42.9	45.5	27.3	42.9	28.6
	4	--	--	--	--	--	--	--	--	50.0	66.7
187	1	87.5	62.5	66.7	83.3	100.0	100.0	27.3	18.2	61.5	61.5
	2	87.5	62.5	66.7	83.3	83.3	83.3	45.5	27.3	69.2	38.5
	3	75.0	50.0	66.7	66.7	83.3	100.0	27.3	27.3	66.7	33.3
	4	--	--	--	--	--	--	--	--	75.0	50.0
187	1	80.0	100.0	50.0	50.0	100.0	85.7	54.5	81.8	100.0	80.0
	2	100.0	100.0	66.7	50.0	100.0	85.7	90.9	90.9	50.0	66.7
	3	80.0	100.0	25.0	0.0	100.0	85.7	72.7	90.9	71.4	57.1
	4	--	--	--	--	--	--	--	--	66.7	83.3
194	1	100.0	100.0					100.0	75.0	100.0	100.0
	2	100.0	100.0					100.0	87.5	100.0	80.0
	3	100.0	92.9					100.0	87.5	100.0	100.0
	4	--	--					--	--	100.0	87.5
196	1	100.0	85.7					87.5	87.5	85.7	85.7
	2	100.0	71.4					87.5	87.5	100.0	100.0
	3	100.0	64.3					100.0	87.5	100.0	100.0
	4	--	--					--	--	100.0	75.0
199	1	87.5	87.5	83.3	100.0	100.0	100.0	81.8	72.7	92.3	100.0
	2	75.0	87.5	83.3	100.0	100.0	83.3	100.0	81.8	84.6	92.3
	3	75.0	87.5	66.7	100.0	100.0	100.0	100.0	81.8	91.7	100.0
	4	--	--	--	--	--	--	--	--	91.7	91.7
209	1	100.0	100.0					62.5	75.0	100.0	100.0
	2	100.0	100.0					75.0	87.5	100.0	100.0
	3	100.0	92.9					75.0	87.5	100.0	100.0
	4	--	--					--	--	100.0	100.0
210	1	78.6	100.0					25.0	75.0	71.4	100.0
	2	71.4	100.0					25.0	87.5	60.0	80.0
	3	57.1	78.6					25.0	87.5	83.3	83.3
	4	--	--					--	--	25.0	87.5
210	1	60.0	60.0	50.0	16.7	71.4	100.0	45.5	81.8	40.0	40.0
	2	60.0	60.0	50.0	33.3	57.1	100.0	63.6	81.8	50.0	33.3
	3	60.0	40.0	0.0	0.0	71.4	100.0	54.5	81.8	71.4	42.9
	4	--	--	--	--	--	--	--	--	--	--

TSE Score	Rating	Nurses		Pharmacists		Physicians		Veterinarians		Consumers	
		Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 2
212	1	100.0	100.0					12.5	87.5	85.7	100.0
	2	85.7	100.0					37.5	87.5	60.0	100.0
	3	92.9	100.0					37.5	100.0	100.0	83.3
	4	--	--					--	--	62.5	87.5
215	1	50.0	62.5	66.7	66.7	83.3	50.0	27.3	45.5	46.2	61.5
	2	50.0	62.5	66.7	66.7	83.3	83.3	27.3	54.5	46.2	69.2
	3	37.5	50.0	66.7	66.7	83.3	50.0	27.3	54.5	50.0	66.7
	4	--	--	--	--	--	--	--	--	41.7	58.3
216	1	60.0	80.0	66.7	83.3	85.7	100.0	81.8	90.9	60.0	100.0
	2	40.0	80.0	66.7	66.7	100.0	100.0	90.9	100.0	66.7	100.0
	3	40.0	80.0	25.0	25.0	100.0	100.0	63.6	90.9	57.1	100.0
	4	--	--	--	--	--	--	--	--	83.3	100.0
224	1	100.0	62.5	83.3	83.3	100.0	83.3	81.8	45.5	76.9	84.6
	2	100.0	62.5	83.3	83.3	100.0	66.7	100.0	81.8	92.3	61.5
	3	100.0	50.0	83.3	66.7	100.0	50.0	90.9	72.7	83.3	50.0
	4	--	--	--	--	--	--	--	--	91.7	66.7
231	1	100.0	80.0	83.3	100.0	100.0	100.0	72.7	81.8	100.0	100.0
	2	100.0	60.0	83.3	100.0	100.0	100.0	90.9	90.9	100.0	100.0
	3	60.0	40.0	50.0	100.0	100.0	100.0	81.8	90.9	100.0	100.0
	4	--	--	--	--	--	--	--	--	100.0	100.0
241	1	87.5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	92.3	100.0
	3	87.5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	91.7	100.0
	4	--	--	--	--	--	--	--	--	100.0	100.0
258	1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	81.8	100.0	92.3
	2	100.0	100.0	100.0	100.0	100.0	100.0	90.9	90.9	100.0	100.0
	3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	90.9	100.0	100.0
	4	--	--	--	--	--	--	--	--	100.0	100.0
268	1	100.0	100.0					100.0	100.0	100.0	100.0
	2	100.0	100.0					100.0	100.0	100.0	100.0
	3	100.0	92.9					100.0	100.0	100.0	100.0
	4	--	--					--	--	100.0	100.0
272	1	100.0	100.0	83.3	50.0	100.0	71.4	63.6	18.2	60.0	80.0
	2	80.0	80.0	83.3	50.0	85.7	85.7	81.8	63.6	66.7	33.3
	3	80.0	80.0	50.0	0.0	85.7	85.7	81.8	45.5	71.4	71.4
	4	--	--	--	--	--	--	--	--	100.0	83.3

TOEFL Research Reports currently available . . .

Report 1. *The Performance of Native Speakers of English on the Test of English as a Foreign Language.* John L. D. Clark. November 1977.

Report 2. *An Evaluation of Alternative Item Formats for Testing English as a Foreign Language.* Lewis W. Pike. June 1979.

Report 3. *The Performance of Non-Native Speakers of English on TOEFL and Verbal Aptitude Tests.* Paul J. Angelis, Spencer S. Swinton, and William R. Cowell. October 1979.

Report 4. *An Exploration of Speaking Proficiency Measures in the TOEFL Context.* John L. D. Clark and Spencer S. Swinton. October 1979.

Report 5. *The Relationship between Scores on the Graduate Management Admission Test and the Test of English as a Foreign Language.* Donald E. Powers. December 1980.

Report 6. *Factor Analysis of the Test of English as a Foreign Language for Several Language Groups.* Donald E. Powers and Spencer S. Swinton. December 1980.

Report 7. *The Test of Spoken English as a Measure of Communicative Ability in English-Medium Instructional Settings.* John L. D. Clark and Spencer S. Swinton. December 1980.

Report 8. *Effects of Item Disclosure on TOEFL Performance.* Gordon A. Hale, Paul J. Angelis, and Lawrence A. Thibodeau. December 1980.

Report 9. *Item Performance Across Native Language Groups on the Test of English as a Foreign Language.* Donald L. Alderman and Paul W. Holland. August 1981.

Report 10. *Language Proficiency as a Moderator Variable in Testing Academic Aptitude.* Donald L. Alderman. November 1981.

Report 11. *A Comparative Analysis of TOEFL Examinee Characteristics, 1977-1979.* Kenneth M. Wilson. September 1982.

Report 12. *GMAT and GRE Aptitude Test Performance in Relation to Primary Language and Scores on TOEFL.* Kenneth M Wilson. October 1982.

Report 13. *The Test of Spoken English as a Measure of Communicative Ability in the Health Professions: Validation and Standard Setting.* Donald E. Powers and Charles W. Stansfield. January 1983.

If you wish additional information about TOEFL research or would like to be placed on the mailing list to automatically receive order forms for newly published reports, write to:

TOEFL Program Office
Educational Testing Service
Princeton, NJ 08541
USA